Dictionnaire des

Mathématiques

algèbre,
analyse
géométrie

PRÉFACE

Le monde scientifique est de plus en plus imprégné de mathématiques, ou de mathématique — le singulier, rare dans le langage courant, semblant néanmoins préférable. C'est pourquoi cette discipline a pris une place de choix dans l'enseignement : tout élève, au cours de ses études, y est nécessairement confronté. Mais, à la fois valorisée et redoutée, elle garde, au-delà des rudiments dispensés au collège et au lycée, une aura de mystère : seuls quelques initiés ont le privilège de faire des recherches en mathématiques, ou même simplement d'avoir une claire vision de ce qu'elles sont.

Dès sa première édition (1968-1974), l'*Encyclopædia Universalis* a voulu offrir au public une vue d'ensemble des mathématiques contemporaines et de leur développement historique. L'ambition de ce projet — les exigences propres à la présentation de cette discipline s'ajoutant à celles qui sont inhérentes à toute entreprise encyclopédique — en rehausse la réussite.

Le présent ouvrage, qui rassemble l'essentiel des questions d'algèbre, analyse, arithmétique et théorie des nombres, géométrie, topologie, algèbre topologique et géométrie algébrique, offre un vaste panorama qui permet de saisir la démarche, les acquis, les avancées des inventeurs de cette architecture abstraite qu'est la mathématique. Un second volume réunira les interrogations sur les fondements, les articles spécifiques historiques, ainsi que tout ce qui touche aux probabilités, aux statistiques et à la plupart des applications.

« Architecture abstraite », disions-nous. En effet, à partir de quelques notions premières, telles qu'« ensemble », « élément », « appartenance », et de quelques axiomes, les structures mathématiques — dans le cadre desquelles tous calculs et démonstrations se font — ne se déploient-elles pas progressivement les unes à partir des autres, des plus « simples » (ensemble ordonné, groupe, espace topologique...) aux plus « subtiles » (espace disqué, espace localement annelé...) ? Là, semble-t-il, réside la beauté mathématique, ou plutôt la partie la plus abstraite de cette beauté car, s'il est de belles théories, il est aussi de jolies formules ($e^{i\pi} = -1$) et la formule de Stirling, par exemple), de beaux calculs, de splendides démonstrations et, bien sûr, de manière plus visible, des courbes dont l'harmonie n'échappe à personne.

Bien entendu, la contemplation de cette beauté exige un minimum de compréhension, jusqu'où il convient de hisser son esprit : songeons qu'en musique ou dans le domaine sportif de sérieux entraînements sont nécessaires si l'on veut trouver vraiment du plaisir à jouer d'un instrument ou à pratiquer un sport, a

fortiori si l'on veut accéder aux concerts ou aux compétitions. Mais, au moins à partir d'un certain niveau, cette activité, sérieuse certes, acquiert une dimension ludique : les mathématiques, pour qui les aime, ouvrent aussi sur tout un espace de jeux. De sorte que la résolution d'un joli problème peut se révéler aussi distrayante que, par exemple, une partie d'échecs ou de shogi (un jeu japonais proche des échecs). Joie de chercher, joie de trouver, joie de la communion enfin avec une beauté qui, pour abstraite qu'elle soit, n'en suscite pas moins de très réels plaisirs : qui est capable de faire ainsi des mathématiques est dans une situation très voisine de celle de l'alpiniste.

Au lecteur, à la lectrice, qu'il ou qu'elle soit ou non mathématicien ou mathématicienne, à tout lecteur tel que le définissait Paul Valéry – c'est-à-dire « de bonne foi » autant que « de mauvaise volonté » – de relever le défi...

Mais la mathématique est aussi et d'abord un langage et, de ce point de vue, intéresse linguistes et lexicographes. Chaque mot ou locution reçoit une définition précise et, à côté de termes spécifiquement mathématiques (morphisme, simplexe...), d'adjectifs honorant un mathématicien (euclidien, eulérien, népérien...) ou une mathématicienne (noethérien...), figurent un assez grand nombre de substantifs (anneau, clan, corps, distribution, fibre, groupe, lacet, spectre, tribu...) ou d'adjectifs (complet, conforme, séparé, simple...) empruntés à la langue courante mais avec un sens mathématique précis, où l'aspect métaphorique est d'ailleurs parfois présent (filtre, noyau, treillis...). De sorte que, au-delà de son aspect faussement ésotérique, il y a parfois une certaine poésie, voire une poésie certaine – osons aller jusque-là ! –, dans le langage mathématique. Avec une pointe d'humour, un grand bol d'enthousiasme et une réserve inépuisable de persévérance, chevauchons donc (sur un paraboloïde hyperbolique, évidemment) à travers les univers mathématiques pour y découvrir les corps algébriquement clos, les endomorphismes diagonalisables, les espaces bornologiques, les fonctions holomorphes ou les produits de convolution.

L'aventure mathématique, commencée sans doute depuis qu'Adam et Ève ont pensé qu'ils étaient deux, n'a certes pas fini de nous passionner. Le présent volume de la collection « Encyclopædia Universalis » nous rappelle – alors que le grand théorème de Fermat vient enfin, après plus de trois siècles de travaux, d'être démontré – qu'il reste bien des questions simples non résolues, par exemple celle-ci : existe-t-il une infinité de nombres premiers « jumeaux », c'est-à-dire de nombres premiers consécutifs dont la différence est deux ?

L'Éditeur

INTRODUCTION

Présenter les mathématiques contemporaines dans le contexte d'une encyclopédie destinée au grand public cultivé pouvait paraître une gageure. Le sujet, dont la réputation d'inaccessibilité n'est plus à faire, paraissait devoir être esquivé. Néanmoins l'*Encyclopædia Universalis* n'a pas hésité, dans ce domaine comme dans tous les autres, à garder le modèle qu'elle s'était donné : l'encyclopédie que Diderot et d'Alembert ont élaborée pour l'époque des Lumières. Considérant qu'il était du devoir des scientifiques de partager leur savoir avec l'ensemble du monde cultivé, Diderot et d'Alembert ont découpé le savoir mathématique qui leur était contemporain en autant de disciplines et de sous-disciplines, et ont alors demandé à ses scientifiques de premier ordre de rédiger ces présentations à caractère introductif.

C'est le même pari qui a été à l'origine de la première conception du traitement des mathématiques dès l'édition de 1968 de l'*Encyclopædia Universalis* : traiter des mathématiques dans leur forme contemporaine de manière à en rendre accessibles les concepts fondamentaux, les principaux courants, les résultats décisifs à des lecteurs possédant une formation scientifique minimale. À ce défi, l'on doit ce qui constitue la présentation la plus complète, la plus approfondie de l'état actuel des mathématiques dans une entreprise encyclopédique de langue française. Une telle tentative n'aurait pu être conçue sans l'intervention décisive de Jean Dieudonné. C'est à lui que nous devons l'élaboration en quelques mois d'un découpage initial des mathématiques qui allait constituer la trame de l'entreprise pour la première édition. Il a par ailleurs participé activement, comme auteur, à de nombreux articles clés avec le style exceptionnel qui le caractérisait et qui a donné le ton à l'ensemble de l'œuvre.

Les mathématiques sont en effet introduites ici de manière historique, dans la mesure où la genèse des concepts contemporains nous a paru constituer le meilleur accès qui puisse y conduire. La présentation de la partie historique des mathématiques se prolonge par l'exposé de ces théories sous leur forme moderne. Le XIXe siècle constitue un moment charnière dans l'évolution des mathématiques, et leur présentation s'articule autour de ce constat. Ainsi, l'algèbre se réduit d'abord à la théorie des équations, et c'est au XIXe siècle que vont naître toutes les nouvelles structures. L'article de synthèse sur l'algèbre renvoie aux différentes entrées du dictionnaire. L'évolution des conceptions de la géométrie est évoquée dans la grande fresque du père Russo. Elle débouche, de manière naturelle, d'une part, sur la géométrie différentielle, d'autre part, sur l'étude des courbes algébriques, point de départ de la géométrie algébrique, qui à son tour rencontre la théorie des nombres.

Fidèle à son impératif d'actualité, l'*Encyclopædia Universalis* a voulu par la suite refléter la transformation qu'avaient subie les mathématiques, tout comme l'ensemble des domaines de la pensée scientifique. La période de 1968 représentait l'apogée du structuralisme et du formalisme, qui s'exprimaient en mathématiques par la prédominance absolue du bourbakisme, lequel marginalisait certaines branches de la discipline. Les décennies ultérieures devaient voir cet impérialisme s'effriter, et d'autres domaines des mathématiques, comme les mathématiques appliquées jusqu'alors absentes, retrouver leur droit de cité. L'édition de 1984 tient compte de cette mutation, grâce à une modification portant sur près de 30 p. 100 des textes.

Des quelque deux cent cinquante articles ou notices publiés dans le Corpus de l'*Encyclopædia Universalis* ainsi que dans les différents volumes annuels ou thématiques, ce *Dictionnaire* reprend uniquement les articles qui présentent les grands problèmes dégagés au XIXe siècle et suivent leurs transformations jusqu'à l'époque contemporaine. Le lecteur pourra s'y familiariser avec les nouveaux objets mathématiques dont l'apparition devait constituer les mathématiques modernes. Il y trouvera une présentation contemporaine des théories classiques. Un autre ensemble d'articles, traitant des questions épistémologiques historiques ou d'aspects plus contemporains fera l'objet d'un recueil ultérieur.

Bien que d'origine ancienne, certaines applications des mathématiques, comme l'analyse numérique, figureront dans ce second volume. Quant au calcul des probabilités, dont la naissance date du XVIIe siècle, sa présentation axiomatique contemporaine, due à Kolmogorov, date de 1933. C'est la raison pour laquelle nous en avons réservé, ainsi qu'à la logique mathématique, l'exposé dans ce volume ultérieur.

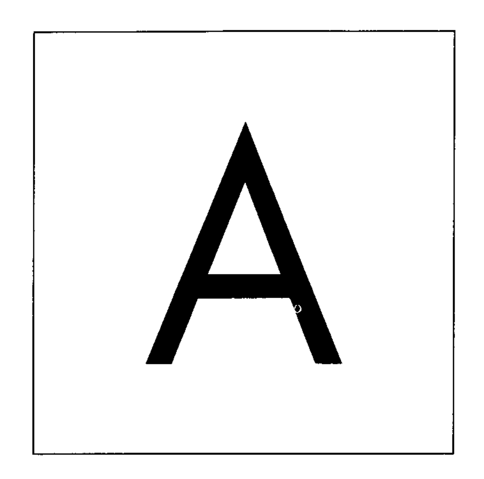
Ce *Dictionnaire des mathématiques* se veut ainsi le reflet du foisonnement et de l'enchevêtrement des diverses disciplines mathématiques. Comme le soulignait souvent Jean Dieudonné, c'est cette « interdisciplinarité » interne qui fait la force et l'originalité des mathématiques contemporaines. Les articles de l'*Encyclopædia Universalis* ont tenté de rendre compte du caractère constamment ouvert de cette démarche.

Jean-Luc VERLEY

COMMENT UTILISER L'INDEX

Placé en fin de volume, c'est l'Index qui donne sa valeur proprement encyclopédique à ce dictionnaire. C'est par lui que toute recherche ou, plus généralement, toute consultation devraient commencer. Nous avons adopté pour sa constitution un certain nombre de conventions qui nous sont propres. Le lecteur les trouvera définies ci-après, exemples à l'appui, sous la forme d'un tableau.

Exemple	Explication
• **BARYCENTRE** *63*	**ENTRÉE précédée d'une puce et suivie d'un numéro de page :** signifie que cette entrée est le titre d'un article du dictionnaire, commençant à la page indiquée
• **HILBERT** ESPACE DE *596* ALGÈBRE *22* ERGODIQUE (THÉORIE) *332* GROUPES - Représentation linéaire des groupes *559* HARMONIQUE (ANALYSE) *587, 592* NORMÉES (ALGÈBRES) *726*	ce même type d'entrée peut être suivi de références
GAUSS CARL FRIEDRICH (1777-1855) ALGÈBRE *14, 17* COMPLEXES (NOMBRES) *116* DIOPHANTIENNES (APPROXIMATIONS) *255* DIOPHANTIENNES (ÉQUATIONS) *263, 267* DIVISIBILITÉ *290*	**ENTRÉE simple** suivie de références
• **SUITES** CALCUL INFINITÉSIMAL - Calcul à une variable *71* CONVEXITÉ - Fonctions convexes *146* DISTRIBUTIONS *276* FONCTIONS (REPRÉSENTATION ET APPROXIMATION DES) *364, 373, 385*	**RÉFÉRENCE à un article long :** titre d'article et numéro de page localisant la partie de texte pertinente au sein de l'article
LIMITE (NOTION DE)	**RÉFÉRENCE à un article court :** titre de l'article
	RENVOIS d'un terme à un autre
NAPIER JOHN ► **NEPER** JOHN	pour des raisons relevant de l'orthographe ou du système de transcription
CALCUL SYMBOLIQUE ► **SYMBOLIQUE** CALCUL	pour des raisons de choix alphabétique
ALEMBERT THÉORÈME DE D' ► **ALGÈBRE** THÉORÈME FONDAMENTAL DE L'	pour des raisons d'ordre sémantique



AFFINE APPLICATION

Soit E et F deux espaces vectoriels sur un corps commutatif K et A et B des espaces affines attachés à E et F. On dit qu'une application u de A dans B est une application linéaire affine (ou application affine) si, quelle que soit la famille finie d'éléments (M_i, λ_i), pour $1 \leqslant i \leqslant k$, où k est quelconque, de A × K, possédant un barycentre G, $u(G)$ est le barycentre des éléments $(u(M_i), \lambda_i)$ de B × K.

On démontre les résultats suivants :

1. Il existe une application linéaire f et une seule de E dans F telle que, pour tout M et tout N dans A et pour $M' = u(M)$ et $N' = u(N)$:

$$f(\overrightarrow{MN}) = \overrightarrow{u(M)u(N)}\,;$$

f s'appelle l'application linéaire associée à u.

2. La composée $v \circ u$ de deux applications affines u et v est une application affine et l'application linéaire associée à $v \circ u$ est $g \circ f$ (où f et g désignent les applications linéaires associées à u et v).

3. Les applications linéaires affines bijectives d'un espace affine A dans lui-même forment un groupe, appelé groupe affine de A et noté **GA**(A). Une application affine u de A dans A est bijective si et seulement si son application linéaire associée f est aussi bijective. Ainsi l'application qui à u fait correspondre f est un morphisme du groupe affine **GA**(A) dans le groupe linéaire **GL**(E).

4. Soit A et B deux espaces affines de dimensions finies (dim A $= q$). Pour définir une application affine de A dans B, il suffit de se donner $(q + 1)$ points affinement indépendants dans A et leurs images dans B.

JACQUES MEYER

AFFINES ESPACE & REPÈRE

Dans la conception intuitive de l'espace usuel, il n'y a pas d'origine privilégiée ; c'est une fois qu'une origine est choisie que cet espace devient un espace vectoriel. La structure d'espace affine formalise cette situation à partir de la notion de translation associée à un vecteur d'extrémités données, défini comme bipoint. Plus précisément, la structure affine se définit comme suit.

Espace affine. Soit E un espace vectoriel sur un corps commutatif K. Un ensemble A est dit espace attaché à l'espace E s'il est muni d'une application de A × E dans A, notée $(M, x) \mapsto M + x$, telle que le groupe additif de E opère simplement transitivement sur A, *i.e.* telle que à $(M, x) \in A \times E$ correspond un point N de A et un seul, tel que $N = M + x$; et à un couple quelconque de points (M, N) de A × A, que l'on désigne sous le nom de bipoint, correspond dans E un vecteur x (appelé opéra-

teur de translation de A) et un seul, tel que $N = M + x$. Ce vecteur x se note $\overrightarrow{MN}$. Deux bipoints AB et CD sont dits équipollents si $\overrightarrow{AB} = \overrightarrow{CD}$.

Soit O un point quelconque de A. Le couple (A, O) s'appelle espace affine muni de l'origine O. L'application de A dans E, définie par $M \mapsto x = \overrightarrow{OM}$, est une bijection qui permet d'identifier l'espace A muni de l'origine O à l'espace vectoriel E.

Réciproquement, par l'application qui à tout couple de vecteurs (x, y) de E associe le vecteur $x + y$, l'ensemble E devient un espace affine attaché à l'espace vectoriel E. Le vecteur nul de E s'appelle origine canonique de l'espace affine E.

Si l'espace E est de dimension finie, on pose $\dim(A) = \dim(E)$.

Variété linéaire affine. Un sous-ensemble $A' \subset A$ est appelé variété linéaire affine (ou variété linéaire) de l'espace affine A si, pour toute famille finie de points de A′, tout barycentre de ces points appartient à A′. Une condition nécessaire et suffisante pour qu'une partie non vide A′ de A soit une variété linéaire affine est que, en prenant un point O quelconque dans A′, l'ensemble des vecteurs $\overrightarrow{OM}$, où $M \in A'$, soit un sous-espace vectoriel E′ de l'espace vectoriel E auquel est attaché A. Le sous-espace E′ ne dépend d'ailleurs pas du choix de O dans A′. D'autre part, on peut montrer que la variété linéaire A′ est un espace affine attaché à E′ (qui est appelé direction de A′). Si E′ est de dimension finie, on pose : $\dim(A') = \dim(E')$. Étant donné un sous-ensemble B de A, on appelle variété linéaire affine engendrée par B la plus petite variété linéaire contenant B ; on montre que c'est l'intersection de toutes les variétés contenant B. D'autre part, la variété linéaire affine engendrée par $(k + 1)$ points de A notés (a_i), pour $1 \leqslant i \leqslant k + 1$, est l'ensemble des barycentres des a_i. Par définition, les $(k + 1)$ points a_i sont dits affinement indépendant (ou forment une famille affinement libre) si la dimension de la variété linéaire qu'ils engendrent est égale à k ; si cette dimension est inférieure à k, ils sont dits affinement liés.

Repère affine. On appelle repère affine d'un espace affine A attaché à un espace vectoriel E de dimension n la donnée d'un point O de A et d'une base $\mathcal{B}$ de E. Le point O est l'origine du repère et les coordonnées d'un point M sont les composantes de $\overrightarrow{OM}$ sur la base $\mathcal{B}$. Ainsi, si :

$$\mathcal{B} = (e_i),$$

pour $1 \leqslant i \leqslant n$, et si :

$$\overrightarrow{OM} = \sum_{i=1}^{n} x_i e_i,$$

les coordonnées de M sont les x_i.

Géométrie affine. La géométrie affine est l'étude des espaces affines et des variétés linéaires affines ainsi que des invariants par le groupe affine.

JACQUES MEYER

ALGÈBRE

L'algèbre au sens moderne, à savoir l'étude des structures algébriques indépendamment de leurs réalisations concrètes, ne s'est dégagée que très progressivement au cours du XIXe siècle, en liaison avec le mouvement général d'axiomatisation de l'ensemble des mathématiques et la préoccupation croissante des mathématiciens de « substituer les idées au calcul » ; jusqu'alors, le propos essentiel de l'algèbre avait été la résolution, par des

formules explicites, des équations algébriques. Les tentatives infructueuses pour résoudre les équations générales de degré supérieur ou égal à cinq, ainsi que les problèmes de la théorie des nombres, conduisirent alors les mathématiciens à introduire des êtres mathématiques de nature nouvelle qui présentaient entre eux des analogies étroites dans leur maniement et par suite à ressentir le besoin de dégager ce qui pouvait être commun à toutes ces situations. Ils furent ainsi amenés à penser que la « nature » des objets mathématiques étudiés est au fond secondaire, et le mathématicien anglais George Boole pouvait déclarer en 1847 : « La mathématique traite les opérations considérées en elles-mêmes, indépendamment des matières diverses auxquelles elles peuvent être appliquées. »

Tout au long du XIXe siècle va se développer ce processus d'axiomatisation de l'algèbre qui aboutit aux structures actuelles. Si, dès 1850, les mathématiciens anglais ont dégagé avec une parfaite netteté la notion de loi de composition et l'appliquent à des situations variées (vecteurs, matrices, algèbre de la logique), il faudra attendre 1910 pour trouver dans la vaste synthèse de Steinitz l'exposé abstrait qui marque le début de l'algèbre moderne proprement dite.

L'étude des groupes domine tout d'abord les préoccupations de cette époque ; introduite par Cauchy et surtout mise en évidence par Galois qui en a montré l'importance dans la théorie des équations, cette notion va jouer un rôle essentiel dans presque tous les domaines des mathématiques, en physique et en mécanique quantique. Les travaux des mathématiciens allemands sur les nombres algébriques seront à l'origine de l'étude des corps et des anneaux commutatifs et ces notions apparaîtront comme les outils essentiels pour étudier les courbes et surfaces algébriques, conduisant à la géométrie algébrique abstraite ; ainsi s'introduit le langage géométrique en algèbre commutative. L'algèbre linéaire prend une grande importance lorsqu'après une axiomatisation convenable les mathématiciens s'aperçoivent du caractère linéaire de nombreuses situations et de l'importance du processus de linéarisation. Et comme « la mathématique est un organisme dont la force vitale a pour condition l'indissoluble union de ses parties » (Hilbert, *Conclusion de la conférence de 1900*), l'algèbre a rejoint avec succès l'analyse par la considération simultanée, sur un même ensemble, de structures algébriques et topologiques (constituant ainsi la branche des mathématiques appelée algèbre topologique).

1. La théorie des groupes

La structure de groupe

La structure de groupe est une des structures algébriques les plus simples et, sans conteste, la plus importante des mathématiques modernes. Son universalité ne s'arrête pas là : le psychologue Piaget a mis en évidence le rôle essentiel joué par cette notion dans les mécanismes mêmes de la pensée, et H. Poincaré a pu dire que la notion de groupe préexiste dans notre esprit car la géométrie ne se concevrait pas sans elle. Cependant, il a fallu presque un siècle pour que se dégage sous forme abstraite cette notion.

Axiomatiquement, un groupe est un ensemble muni d'une loi de composition

interne $(x, y) \mapsto x*y$ associative [c'est-à-dire $(x*y)*z = x*(y*z)$] telle qu'il existe un élément privilégié e, appelé élément neutre, tel que $x*e = e*x = x$ et telle que tout élément ait un inverse (c'est-à-dire pour tout x il existe un élément y tel que $x*y = y*x = e$). Un tel groupe est dit abélien, ou commutatif, si $x*y = y*x$.

Les ensembles usuels de nombres (entiers relatifs, nombres rationnels, nombres complexes) sont des groupes abéliens pour l'addition ; les ensembles des nombres rationnels non nuls, ou réels non nuls, sont des groupes abéliens pour la multiplication. Un important exemple de groupe non commutatif est celui des transformations de notre espace usuel à trois dimensions qui conservent la distance de deux points (ce sont les déplacements). Elles constituent un groupe non abélien si on convient que le produit $S * T$ de deux transformations S et T est la transformation obtenue en effectuant successivement la transformation T puis la transformation S.

Les groupes finis

Le premier exemple de groupe formé d'éléments de nature assez différente de celle des nombres est fourni par les travaux de Gauss sur les formes quadratiques $ax^2 + bxy + cy^2$, où a, b, c sont des entiers relatifs premiers entre eux. Deux telles formes étant dites équivalentes si l'on passe de l'une à l'autre par un changement de variable $x' = px + qy$ et $y' = rx + sy$, où p, q, r, s sont des entiers relatifs tels que $ps - qr = 1$, Gauss définit sur l'ensemble des classes de formes, de discriminant $D = b^2 - 4\,ac$ donné, une loi de composition qui en fait un groupe abélien fini. Dans ses *Disquisitiones arithmeticae* de 1801, Gauss rencontre également d'autres groupes finis tels que le groupe additif des entiers modulo un entier m ou le groupe multiplicatif des racines m-ièmes de l'unité dans le corps des nombres complexes, mais la notion de groupe n'apparaît pas formulée avec netteté avant Cauchy. En 1830, dans ses travaux sur la résolubilité des équations algébriques, Galois ramène l'étude d'une telle équation à celle du groupe (fini) de permutations de ses racines ; à ce propos, l'auteur introduit les notions fondamentales de sous-groupe distingué et de suite normale. Les groupes finis, et plus précisément les groupes de permutations, vont être l'objet presque exclusif de la théorie des groupes pendant de nombreuses années ; les résultats les plus profonds obtenus dans ce domaine au XIXe siècle sont ceux de Jordan (*Traité des substitutions et des équations algébriques*, Paris, 1870) et de Sylow sur la structure des groupes finis. Beaucoup plus récemment, en liaison avec des préoccupations d'arithmétique et de géométrie algébrique, la théorie des groupes finis a connu un nouvel essor ; les découvertes les plus spectaculaires de ces dernières années sont surtout relatives aux caractères et aux représentations linéaires de ces groupes : travaux de Brauer, Chevalley, Feit-Thomson, Novikov (cf. GROUPES FINIS et représentation linéaire des GROUPES).

Groupes et géométrie

C'est à Jordan que remonte la première étude de groupes contenant une infinité d'éléments, notion qui allait prendre une importance considérable durant la deuxième moitié du XIXe siècle. En liaison avec le renouveau des études géométriques et les préoccupations axiomatiques de cette époque, la notion de groupe de transformation va prendre un essor considérable avec l'étude systématique des invariants d'un tel groupe, *i.e.* l'étude des propriétés

qui ne sont pas modifiées par les transformations du groupe. Ainsi, dans notre espace usuel à trois dimensions, les angles et les distances ne sont pas changés par un déplacement, les angles et les rapports de longueurs restent invariants par une similitude, la notion de parallélisme ou la nature d'une conique sont invariantes par une transformation linéaire régulière des coordonnées. C'est F. Klein, dans son célèbre « programme d'Erlangen », de 1872, qui dégagera un principe général, que nous énoncerons sous une forme volontairement vague et intuitive : la donnée d'un espace et d'un groupe de transformations opérant sur cet espace définit une « géométrie », qui est l'étude des propriétés qui restent invariantes lorsqu'on applique les transformations du groupe. Ainsi, la géométrie métrique (resp. affine, resp. projective) est l'étude des propriétés invariantes par le groupe orthogonal (resp. affine, resp. projectif) et cette théorie constitue un langage commun qui englobe à la fois les géométries euclidiennes et non euclidiennes construites à cette époque (cf. GROUPES CLASSIQUES, GÉOMÉTRIE) ; la théorie de la relativité allait attirer l'attention sur la géométrie construite à partir du groupe de Lorentz, qui joue un rôle essentiel dans les théories quantiques.

Les travaux de Klein allaient également mettre en évidence la notion des groupes isomorphes : en 1877, Klein découvre que le groupe de permutation des racines de l'équation du cinquième degré est substantiellement identique au groupe des transformations du polyèdre régulier appelé icosaèdre ; bien que techniquement cette notion de groupes isomorphes ait été utilisée par Galois et même Gauss dans des cas particuliers, elle n'apparaît sous forme générale qu'à cette époque. En fait, ce n'est que beaucoup plus récemment que la notion d'isomorphisme a pris toute sa valeur, avec les développements de l'axiomatique mettant en évidence le fait que toute structure porte en elle une notion d'isomorphisme. Cette « identification » des groupes isomorphes allait conduire à la théorie de la représentation linéaire des groupes, qui est la recherche et l'étude de groupes de matrices isomorphes (ou, à défaut, homomorphes) à un groupe donné.

Les travaux précédents sur la géométrie avaient mis en évidence l'importance des « groupes continus » ; sous l'action de S. Lie et de ses élèves, puis de É. Cartan, cette notion allait être le germe d'une des théories les plus centrales des mathématiques contemporaines : la théorie des groupes de Lie (cf. GROUPES - Groupes de Lie), tandis que l'exemple des groupes classiques conduisait à la théorie des groupes algébriques qui admet d'importantes applications en géométrie algébrique et en théorie moderne des nombres.

2. Les origines de l'algèbre commutative

Corps et anneaux

L'étude des corps et des anneaux trouve son origine dans les travaux de l'école allemande du XIX[e] siècle, principalement ceux de Kummer, Kronecker, Dedekind et Hilbert. Au départ, les motivations sont ici essentiellement la théorie des équations puis la théorie arithmétique des nombres algébriques, qui découle de recherches relatives au théorème de Fermat ; plus tardivement, et jusqu'à l'époque contemporaine, la géométrie algébrique a été également une source d'idées essentielles.

La notion d'*anneau* dégage sous forme abstraite les analogies constatées par exem-

ple dans le maniement des nombres entiers relatifs et des polynômes : un anneau est un ensemble muni de deux lois de composition internes :

$$(x,y) \mapsto x + y \text{ et } (x,y) \mapsto xy,$$

appelées addition et multiplication respectivement, telles que la première soit une loi de groupe abélien et que la seconde soit associative (*i.e.* $(xy)z = x(yz)$) ; on impose de plus les conditions suivantes de distributivité entre les deux lois :

$$x(y + z) = xy + xz \text{ et } (y + z)x = yx + zx,$$

pour x, y, z quelconques dans l'anneau. Il est commode de supposer l'existence d'un élément unité pour la multiplication. Lorsque, comme dans le cas des nombres rationnels par exemple, l'ensemble des éléments distincts de l'élément neutre pour la première loi (noté 0) est un groupe pour la seconde loi, on dit que l'anneau est un *corps*. Ici on considérera seulement le cas où la multiplication est commutative, en renvoyant à la fin du chapitre 3 le cas non commutatif.

La théorie des corps

Les premiers exemples de corps non triviaux ont été introduits par la théorie des équations. Les travaux de Gauss avaient familiarisé les mathématiciens avec le maniement des nombres complexes et Abel, puis Galois, dégagent l'idée d'adjonction : ils considèrent les corps engendrés par les racines ou les coefficients (indéterminés) d'une équation mais, en fait, si ces auteurs définissent avec précision l'appartenance d'une quantité à un tel corps, ils ne considèrent pas explicitement l'ensemble ainsi constitué. Il faut attendre Dedekind (qui introduit le mot corps) pour une étude systématique de certains corps d'un type assez général, les corps de nombres algébriques ; ce sont des corps $\mathbf{Q}(\theta)$ obtenus de la façon suivante : si θ est un nombre complexe racine d'une équation $f(x) = 0$ de degré n, à coefficients entiers, irréductible sur le corps $\mathbf{Q}$ des nombres rationnels, on appelle $\mathbf{Q}(\theta)$ l'ensemble, qui est un corps, des nombres complexes $a_0 + a_1\theta + ... + a_{n-1}\theta^{n-1}$ où les a_i sont des nombres rationnels quelconques.

Tous les corps de nombres algébriques sont des sous-corps du corps des nombres complexes ; reprenant une idée de Cauchy qui définissait les nombres complexes comme classes résiduelles de polynômes à coefficients réels modulo le polynôme x^2+1, Kronecker donne, en 1882, les premiers exemples de corps (non triviaux) définis abstraitement en montrant que, avec les notations ci-dessus, le corps $\mathbf{Q}(\theta)$ est isomorphe au corps des classes résiduelles de polynômes à coefficients rationnels modulo le polynôme $f(x)$. Vers la même époque, Dedekind et Weber font rentrer dans la théorie des corps le calcul des congruences modulo un nombre premier (mettant ainsi en évidence les premiers corps finis, déjà étudiés par Galois) et donnent une première esquisse d'une théorie axiomatique des corps.

À la fin du XIXe siècle, les exemples de corps définis abstraitement vont se multiplier. Il faut citer surtout les corps de nombres p-adiques, introduits par Hensel et dont l'importance dans de nombreuses branches des mathématiques est considérable, et les corps de séries formelles, introduits par Véronèse en liaison avec des préoccupations de géométrie algébrique. Tous ces exemples allaient conduire Steinitz, en 1910, à développer systématiquement la théorie des corps et de leurs extensions sous la forme qu'elle possède actuellement.

La théorie des idéaux

À l'origine de la théorie des anneaux, on trouve essentiellement des recherches de théorie des nombres. En 1831, Gauss avait été amené, à propos de ses célèbres recherches sur les résidus biquadratiques, à étudier des propriétés de divisibilité dans l'anneau $\mathbf{Z}[i]$ des « entiers de Gauss » de la forme $a + bi$, a et b entiers relatifs et $i^2 = -1$; il avait constaté une parfaite analogie avec les propriétés correspondantes de l'anneau $\mathbf{Z}$ des entiers rationnels, ce qui s'explique, dans le langage moderne, par le fait que ces deux anneaux sont principaux. Les travaux de Kummer sur le grand théorème de Fermat allaient faire apparaître des anneaux pour lesquels la situation est souvent très différente ; il s'agit des anneaux cyclotomiques ainsi définis : p étant un nombre premier et ζ étant une racine primitive p-ième de l'unité, on appelle $\mathbf{Z}[\zeta]$ l'ensemble, qui forme un anneau, des combinaisons linéaires à coefficients entiers de puissances de ζ. Comme on le sait, le théorème de Fermat affirme que la relation :

$$x^p + y^p = z^p$$

est impossible pour x, y, z entiers non nuls et p entier supérieur ou égal à trois ; en fait, on voit facilement qu'on peut se borner à établir cette impossibilité pour p premier > 3. Il est probable que la « démonstration » de Fermat utilisait implicitement le fait, erroné dans le cas général, que, comme dans l'anneau $\mathbf{Z}$, tout élément de l'anneau $\mathbf{Z}[\zeta]$ s'écrit de manière *unique* (à un élément inversible dans l'anneau près) comme produit d'éléments premiers. En 1845, après huit ans d'efforts, Kummer en introduisant ses « nombres idéaux » (qu'on appelle maintenant des diviseurs) élucide complètement le problème de la division dans les anneaux cyclotomiques et démontre le théorème de Fermat dans de très nombreux cas.

L'idée de Kummer est en gros la suivante : soit x un élément de l'anneau $\mathbf{Z}[\zeta]$ qui admet deux décompositions différentes, soit, pour simplifier :

$$x = x_1 x_2 \quad \text{et} \quad x = x_3 x_4$$

où les éléments x_1, x_2, x_3, x_4 sont tous premiers ; on suppose de plus que ces deux décompositions ne diffèrent pas seulement par un élément inversible. Kummer démontre qu'on peut représenter les éléments non nuls de l'anneau considéré comme objets d'un nouvel ensemble muni d'une multiplication et dans lequel la décomposition en facteurs premiers est cette fois définie de manière unique. Ainsi, pour tout élément $x \neq 0$ de $\mathbf{Z}[\zeta]$, son image (x), sera décomposable de manière unique en facteurs premiers « idéaux », mais ces facteurs premiers ne sont pas nécessairement les images de certains éléments de l'anneau $\mathbf{Z}[\zeta]$; de même, un élément premier de l'anneau $\mathbf{Z}[\zeta]$ n'a pas nécessairement pour image un « nombre idéal » premier. L'existence de deux décompositions distinctes rencontrées ci-dessus pour x s'explique ainsi : il existe des « nombres idéaux » p_1, p_2, p_3, p_4 tels que :

$$(x_1) = p_1 p_2, \qquad (x_2) = p_3 p_4,$$
$$(x_3) = p_1 p_3, \qquad (x_4) = p_2 p_4,$$

et les deux décompositions de x s'écrivent :

$$(x) = p_1 p_2 \cdot p_3 p_4 = p_1 p_3 \cdot p_2 p_4,$$

qui diffèrent seulement par l'ordre des facteurs.

La notion d'idéal d'un anneau, sous groupe additif qui est stable par multiplication par un élément quelconque de l'anneau, a été introduite, en liaison avec

les travaux de Kummer, par Dedekind dans le cas des anneaux d'entiers algébriques (cf. *infra*). Dedekind montra que les « nombres idéaux » peuvent être représentés par les idéaux de l'anneau, donnant ainsi un exemple de loi de composition entre ensembles d'éléments. En général, un idéal n'est pas inversible pour la loi de composition ainsi définie ; par symétrisation de cette loi, on introduit les idéaux fractionnaires qui sont importants en théorie des nombres et en géométrie algébrique.

Les anneaux auxquels on peut généraliser la théorie de Kummer ont été étudiés systématiquement à l'époque contemporaine, conduisant à la notion générale d'anneau de Dedekind. Un outil essentiel est ici la notion de valuation d'un corps introduite sous forme générale par Krull en 1931 mais déjà utilisée antérieurement dans des cas particuliers, par Ostrowski notamment ; les idéaux premiers d'un anneau de Dedekind sont en correspondance biunivoque avec les classes de valuations équivalentes du corps des fractions de cet anneau.

Éléments entiers

L'étude arithmétique systématique des corps de nombres algébriques n'était possible qu'en introduisant une notion d'élément entier jouant, pour un tel corps, le même rôle que les entiers usuels pour le corps des nombres rationnels. Les progrès dans ce domaine furent réalisés à peu près simultanément et indépendamment par Kronecker et Dedekind pendant la seconde moitié du XIXe siècle. La notion d'*entier algébrique* est due à Dedekind : un nombre complexe est un entier algébrique s'il est racine d'un polynôme à coefficients entiers rationnels dont le coefficient du terme dominant est égal à 1 ; les entiers algébriques d'un corps K de nombres algébriques forment un anneau, que Dedekind appelle un ordre (le mot anneau est de Hilbert). Dans un théorème célèbre et profond, Dirichlet décrit complètement le groupe multiplicatif des éléments inversibles de l'anneau des entiers d'un corps de nombres algébriques et ce résultat a d'importantes applications arithmétiques, notamment dans l'étude des représentations des nombres entiers par des formes quadratiques.

Plus généralement, si A est un anneau contenu dans un corps K, on peut définir les éléments du corps qui sont entiers sur A ; un tel anneau A est dit « intégralement clos » s'il est égal à l'ensemble des éléments de son corps des fractions qui sont entiers sur lui. Ces anneaux ont pris une grande importance en géométrie algébrique contemporaine depuis que Zariski et ses élèves ont mis en évidence l'intérêt des variétés algébriques dites normales, qui possèdent la propriété qu'en chacun de leurs points l'anneau des fonctions rationnelles définies en ce point est intégralement clos.

Géométrie algébrique et algèbre commutative

Il n'est pas question même d'esquisser ici l'histoire de la géométrie algébrique, qui était au départ l'étude des courbes algébriques, et qui, sous sa forme actuelle, la théorie des schémas, due au mathématicien français A. Grothendieck, est devenue une des branches les plus abstraites et les plus vivantes des mathématiques contemporaines ; nous essayerons seulement de montrer, de manière d'ailleurs bien incomplète, comment les premiers besoins de cette science ont conduit à l'introduction et à l'étude axiomatique de nouveaux types d'anneaux.

À l'origine, le propos de la géométrie algébrique était essentiellement l'étude des courbes dans le plan projectif complexe et la théorie des « fonctions algébriques », développée par Weierstrass et Riemann à partir des travaux d'Abel et Jacobi, utilisait presque uniquement des méthodes transcendantes. Avec Riemann et Dedekind, le centre d'intérêt se porte sur l'anneau des fonctions rationnelles partout définies (sauf à l'infini) ; les mathématiciens découvrent alors que les propriétés géométriques de la courbe se reflètent dans cet anneau et que l'étude de ces anneaux et l'étude géométrique vont de pair.

Hilbert, dans ses travaux sur les anneaux de polynômes à plusieurs variables, dégage le fait important que tous les idéaux de ces anneaux sont engendrés par un nombre fini d'éléments. Ces conditions de *finitude* allaient prendre un grande importance avec les travaux de la mathématicienne allemande E. Noether qui, vers 1920, étudie systématiquement ces anneaux (appelés actuellement anneaux noethériens). La géométrie algébrique s'étant progressivement débarrassée, pendant la première moitié du XXe siècle, de toute hypothèse sur le corps de base et la nature des singularités, on peut dire, depuis 1940 environ, que tout résultat relatif aux anneaux noethériens a une interprétation « géométrique » dans ce cadre. Dans cet ordre d'idée, signalons par exemple un résultat important : bien que la théorie de Kummer ne soit pas valable pour l'anneau des polynômes à n variables sur un corps K, on peut cependant associer à tout idéal de cet anneau un ensemble bien déterminé d'idéaux premiers et ceux de ces idéaux premiers qui sont minimaux correspondent aux composantes irréductibles de l'ensemble défini dans K^n par l'annulation des polynômes de l'idéal. Par ailleurs, on peut ici encore donner des théorèmes de « décomposition » des idéaux en introduisant des notions nouvelles qui dépasseraient le cadre de cet article (« décomposition primaire » ; cf. ANNEAUX COMMUTATIFS). Signalons pour terminer que la notion de dimension, directement issue de la géométrie algébrique, a été convenablement axiomatisée pour des anneaux commutatifs très généraux et est étudiée de manière abstraite dans ces anneaux.

Anneaux locaux et localisation

L'anneau $\mathbf{Z}_{(p)}$ des nombres rationnels dont le dénominateur n'est pas divisible par un nombre premier p, ou l'anneau des germes de fonctions holomorphes dans un voisinage de l'origine du plan complexe, possèdent une propriété commune : il existe un idéal de cet anneau qui est distinct de l'anneau et qui contient tous les autres idéaux distincts de l'anneau (dans le premier cas, c'est l'ensemble des nombres rationnels dont le numérateur est divisible par p sans que le dénominateur le soit et, dans le second cas, l'ensemble des germes des fonctions considérées qui s'annulent à l'origine). De manière générale, on appelle anneau local tout anneau possédant cette propriété, et on étudie ces anneaux sous forme abstraite ; l'intérêt de cette notion est qu'elle inclut en particulier tous les anneaux de germes de fonctions (rationnelles, différentiables ou analytiques) que l'on rencontre dans la théorie des variétés algébriques, différentiables ou analytiques. En liaison avec la notion de valuation dont une des applications a déjà été signalée ci-dessus, un autre exemple important d'anneau local est constitué par les anneaux de valuation : un sous-anneau A d'un corps K, qui est distinct de K, est appelé un anneau de valuation de K si, pour tout $x \neq 0$ qui n'appartient pas à A,

son inverse x appartient à A ; ces anneaux correspondent à l'ensemble des éléments de K où une valuation de K prend des valeurs supérieures à 1.

Reprenons l'exemple de l'anneau $\mathbf{Z}_{(p)}$ ci-dessus pour expliquer dans un cas particulier la méthode générale de localisation. Considérons une équation diophantienne :

$$P(x_1, ..., x_n) = 0,$$

où P est un polynôme à coefficients entiers rationnels. Pour trouver les solutions entières de cette équation, on peut d'abord chercher les solutions qui appartiennent au corps des quotients Q de l'anneau **Z**, puis, dans une seconde étape, les solutions rationnelles dont le dénominateur n'est pas divisible par un nombre premier p, *i. e.* les solutions qui appartiennent à l'anneau $\mathbf{Z}_{(p)}$, appelé l'anneau local de **Z** qui correspond au nombre premier p. Bien entendu, si l'équation considérée à une solution dans **Z**, cette solution appartiendra à tous les anneaux locaux $\mathbf{Z}_{(p)}$. Dans le cas d'un anneau général A, on peut de même résoudre le problème posé dans les anneaux locaux correspondant aux idéaux premiers de l'anneau. La résolubilité de l'équation dans chacun des anneaux locaux (localisation) est une condition nécessaire d'existence d'une solution dans l'anneau A. L'étude de la suffisance de ces conditions (en nombre infini dans le cas général) s'appelle la globalisation ; signalons tout de suite qu'en général la globalisation n'est pas possible sous la forme indiquée ci-dessus.

3. L'algèbre linéaire et les origines de l'algèbre non commutative

Structures linéaires

L'étude des équations et systèmes d'équations du premier degré était reléguée au début du XIXe siècle dans l'enseignement élémentaire et négligée des mathématiciens, lorsqu'une axiomatisation convenable montra la puissance des notions nouvelles ainsi mises en évidence. Sous sa forme actuelle, l'algèbre linéaire est une remarquable synthèse conduisant à un vocabulaire et à des résultats qui s'appliquent presque universellement dans tous les domaines des mathématiques et de la physique contemporaine, tandis que le processus de « linéarisation » apparaît comme essentiel dans de nombreuses branches des mathématiques pures et appliquées. La notion fondamentale est ici celle d'espace vectoriel ; elle généralise les propriétés de l'ensemble des vecteurs de notre espace à trois dimensions. Un *espace vectoriel* E sur un corps K est un ensemble d'éléments, appelés « vecteurs », muni d'une loi de groupe abélien notée additivement et d'une loi externe qui à tout couple (a, x) d'un élément a du corps K et d'un vecteur x de E fait correspondre un vecteur $a.x$ de E de telle sorte que l'on ait :

$a.(b.x) = (ab).x$, pour a, b dans K et x dans E ; $1.x = x$, pour tout x de E (1 est l'élément neutre de K pour la multiplication) ;

$(a + b).x = a.x + b.x$, pour a, b dans K et x dans E ;

$a.(x + y) = a.x + b.y$, pour a dans K et x, y dans E.

Les applications d'un tel espace vectoriel E dans un autre qui respectent la structure d'espace vectoriel, *i.e.* telles que :

$$f(x + y) = f(x) + f(y), f(a . x) = a.f(x),$$

pour a dans K et x, y dans E, sont dites *linéaires*.

Une *algèbre* E sur un corps K est un K-espace vectoriel E muni d'un « produit » qui est une loi $E \times E \rightarrow E$ linéaire par rapport à chaque facteur (on dit bilinéaire). Si cette loi est associative,

et admet un élément unité, on a une structure d'anneau.

Par exemple, les nombres complexes forment une algèbre sur le corps des nombres réels.

Espaces de dimension finie

La représentation géométrique des nombres complexes introduite par Argand l'avait amené implicitement à définir l'addition des vecteurs du plan ; plus généralement, la nécessité d'un calcul de nature « géométrique », ou « intrinsèque » (*i.e.* indépendant du choix du système d'axes de coordonnées), allait conduire Grassmann, Möbius et Hamilton à dégager durant la première moitié du XIXe siècle, les règles du calcul vectoriel et, presque simultanément, à généraliser les propriétés de l'espace « usuel » à deux ou trois dimensions en introduisant des espaces de dimension supérieure. Ces derniers apparaissent tout d'abord comme un langage géométrique commode pour interpréter des résultats algébriques valables sans modification pour un nombre quelconque de variables et susceptibles d'une interprétation géométrique dans le cas de deux ou trois variables. Grassmann définit, de manière déjà presque axiomatique, les espaces à n dimensions, l'addition des vecteurs, l'indépendance d'un système de vecteurs, étudie la dimension des sous-espaces vectoriels, sans recours aux coordonnées, et construit l'algèbre extérieure d'un espace vectoriel. Dans ce cadre allait s'insérer tout naturellement l'étude générale des systèmes d'équations linéaires : la notion de rang d'un tel système est dégagée par Frobenius et les résultats généraux sont obtenus par Kronecker ; en liaison avec ces préoccupations, Kronecker et Weierstrass donneront une définition axiomatique des déterminants, déjà connus depuis le XVIIIe siècle et que Grassmann avait rattachés à son calcul extérieur. Les concepts généraux d'algèbre linéaire et multilinéaire relatifs aux espaces vectoriels de dimension finie sont précisés rapidement et on assiste successivement à l'élaboration du calcul matriciel par Cayley et à l'introduction du produit tensoriel par Kronecker ; cependant tous les travaux des mathématiciens de cette époque restent truffés d'hallucinants calculs où les déterminants jouent un rôle essentiel et le caractère intrinsèque des éléments qui interviennent est souvent peu visible.

En liaison avec le renouveau de la géométrie, la notion de dualité se dégage peu à peu pour les espaces vectoriels, mettant en évidence la notion de variables « cogrédientes » ou « contragrédientes », c'est-à-dire variant dans un espace vectoriel ou dans l'espace vectoriel dual. L'étude des coniques et des quadriques, ainsi que de nombreuses recherches arithmétiques avaient mis en vedette les formes quadratiques à 2, 3 puis n variables et les formes bilinéaires qui leur sont associées ; la théorie des invariants, créée par Cayley, Hermite et Sylvester, introduit systématiquement des formes multilinéaires à plusieurs séries de variables cogrédientes et contragrédientes, ce qui, aux notations près, revient à définir des tenseurs. En liaison avec la géométrie différentielle, ces travaux allaient conduire, au début du XXe siècle, Ricci et Levi-Civita à construire le calcul tensoriel et Poincaré et É. Cartan le calcul différentiel extérieur, issu directement de la multiplication extérieure de Grassmann.

Axiomatisation de l'algèbre linéaire

Dès 1888, Peano avait donné une définition axiomatique des espaces vectoriels

généraux (sur le corps des nombres réels) et des applications linéaires dans ces espaces, mais c'est l'analyse qui fournit les plus importants exemples d'espaces vectoriels de dimension infinie et conduisit à saisir toute la portée de l'algèbre linéaire. À propos de recherches sur les équations différentielles et surtout les équations aux dérivées partielles, Hilbert introduit, à l'aube du XXe siècle, le célèbre espace de Schmidt et utilise systématiquement des techniques linéaires pour étudier les opérateurs dans cet espace (cf. espace de HILBERT) et c'est Toeplitz, élève de Hilbert, qui donne la définition d'un espace vectoriel sur un corps quelconque et étend à ces espaces de nombreux résultats d'algèbre linéaire en constatant qu'ils sont indépendants de la théorie des déterminants et subsistent sans supposer que l'espace est de dimension finie. Quelques années plus tard, Banach allait étudier systématiquement les opérateurs linéaires et la dualité dans les espaces vectoriels de fonctions (cf. § 4), tandis qu'Artin, E. Noether et Krull allaient mettre en évidence le caractère presque entièrement linéaire de l'algèbre moderne.

À l'époque contemporaine, on s'est aperçu de l'intérêt qu'il y avait à généraliser la notion d'espace vectoriel en remplaçant le corps de base par un anneau (pas nécessairement commutatif), définissant ainsi la notion de module (dans le cas non commutatif, il faut distinguer des modules à droite et à gauche). L'étude des modules sous forme abstraite s'est épanouie sous l'influence de S. MacLane, H. Cartan et S. Eilenberg pour aboutir à une branche nouvelle de l'algèbre, l'algèbre homologique, issue directement des problèmes posés par la topologie algébrique, et dont le but principal est l'étude des questions où interviennent des relations de dépendance linéaire entre éléments d'un module. L'algèbre homologique est non seulement devenue l'outil essentiel de la topologie algébrique mais est venue féconder de nombreux autres secteurs des mathématiques contemporaines ; sous la forme axiomatique que lui ont donnée H. Cartan et Eilenberg, cette science est également à l'origine de la théorie des catégories abéliennes.

Algèbres non commutatives

Les débuts de l'algèbre non commutative apparaissent étroitement liés à l'élaboration de l'algèbre linéaire. Lorsque Hamilton considéra un nombre complexe $a + bi$ comme un couple ordonné (a, b) de nombres réels, les opérations d'addition et de multiplication entre de tels couples étant celles qui sont déduites du calcul usuel sur les nombres complexes, il chercha en définissant une multiplication convenable, à étendre les propriétés du corps des nombres complexes à des « ternes » ou « quaternes », c'est-à-dire à des systèmes de trois ou quatre nombres réels, pour construire une algèbre jouant pour les rotations de l'espace à trois dimensions le même rôle que les nombres complexes pour les rotations planes de centre O. C'est ainsi qu'il construisit, vers 1845, les quaternions, premier exemple de corps dont la multiplication n'est pas commutative. C'est en essayant de généraliser sa découverte, en introduisant par exemple les biquaternions, que Hamilton fut amené à dégager le fait qu'on peut définir une structure d'algèbre sur un espace vectoriel de dimension finie en donnant la « table de multiplication » des éléments d'une base. Les matrices allaient donner de nombreux autres exemples d'algèbres non commutatives, mais ce n'est qu'en 1870 que B. Pierce donne une définition axiomati-

que et introduit les notions fondamentales relatives aux algèbres de dimension finie ; d'autre part, les travaux de Lie (et de son école) à propos des algèbres qui portent son nom allaient dégager la notion fondamentale de radical et É. Cartan allait mettre en évidence le rôle essentiel joué par les algèbres semi-simples. Une autre application très importante de l'algèbre non commutative est la représentation linéaire des groupes et algèbres, développée de 1896 à 1910 par Frobenius, Burnside et Schur.

4. Algèbre topologique

La continuité des opérations algébriques est d'usage courant dans l'analyse classique ; depuis le début du XIXe siècle, en liaison avec l'introduction des nouveaux êtres mathématiques considérés plus haut, les mathématiciens allaient rencontrer dans de nombreux problèmes de nature variée des ensembles munis d'une notion de convergence et de lois de composition « continues pour cette notion de convergence ». Ce mariage fréquent de l'algèbre et de la topologie a conduit à étudier axiomatiquement ces situations, introduisant ainsi de nouvelles structures très utiles et très riches, qui jouent un rôle essentiel dans de nombreuses théories mathématiques contemporaines ; à titre d'exemple, on peut signaler les espaces vectoriels topologiques et les groupes topologiques.

Espaces vectoriels normés et espaces vectoriels topologiques

Un espace vectoriel normé sur le corps K des nombres réels ou des nombres complexes est un espace vectoriel E sur lequel est définie une fonction $x \longrightarrow \|x\|$, à valeurs réelles positives, possédant les propriétés suivantes, qui généralisent celle de la longueur d'un vecteur dans les espaces de dimension finie :

a) $\|x\| = 0$ si et seulement si $x = 0$;

b) $\|x + y\| \leqslant \|x\| + \|y\|$, pour x, y quelconques dans E ;

c) $\|a.x\| = |a|\|x\|$, pour a dans K et x dans E. ($|a|$ est ici la valeur absolue ou le module du nombre réel ou complexe a).

La considération d'espaces « fonctionnels » (c'est-à-dire d'espaces vectoriels dont les éléments sont des fonctions) munis d'une norme convenable est devenue un des outils essentiel de l'analyse contemporaine.

La théorie des espaces vectoriels normés s'est constituée de 1900 à 1930 approximativement et ici encore l'espace de Hilbert a joué un rôle historique considérable. Hilbert, au début du XXe siècle, fut amené à introduire deux notions de convergence différentes sur l'espace des suites (x_n) de nombres réels tels que la série $\sum_n x_n^2$ soit convergente et étudie la continuité de nombreuses applications linéaires. Quelques années plus tard, vers 1907-1908, Schmidt, Fréchet et Riesz généralisent le langage de la géométrie des espaces de dimension finie à l'espace de Hilbert et introduisent la norme dans ce cas particulier ; la notion d'espace vectoriel normé général apparaît alors vers 1920 dans les travaux de Hahn et de Banach.

Un des aspects essentiels des problèmes sur les espaces vectoriels normés est la théorie de la dualité topologique qui occupe déjà une place centrale dans les travaux de F. Riesz sur les espaces de fonctions intégrables. À partir de 1927, Hahn et Banach abordent de manière générale le problème de la dualité en montrant qu'on peut munir le dual d'un

espace normé d'une structure d'espace normé (complet) ; itérant cette construction, Hahn pourra poser de manière générale le problème des espaces réflexifs, *i. e.* qui sont isomorphes à leur bidual topologique. Vers 1932, la théorie des espaces normés est à peu près achevée avec le livre de Banach, *Théorie des opérations linéaires*.

Une notion telle que la convergence *simple* d'une suite de fonctions dans un espace fonctionnel n'est pas associée à une norme, et il était nécessaire de considérer sur des espaces vectoriels des notions de convergence plus générales que celles définies par des normes, situation étudiée pour la première fois par Fréchet. Mais, sans hypothèse restrictive, la théorie générale était trop pauvre ; la notion essentielle qui allait permettre à la théorie de s'épanouir est la convexité, étudiée par Banach et ses élèves, conduisant von Neumann en 1935 à définir les espaces localement convexes. Des branches essentielles des mathématiques contemporaines, la théorie des distributions par exemple, utilisent de manière constante la théorie de ces espaces.

Groupes topologiques

La nécessité d'étudier des groupes « continus » plus généraux que les groupes de Lie conduisit Schreier en 1927 à définir des groupes dits topologiques, tels que la multiplication et le passage à l'inverse soient des opérations continues. Ceux de ces groupes qui, comme les groupes de Lie, sont localement compacts possèdent des propriétés remarquables dont l'étude constitue une branche nouvelle de l'analyse, l'analyse harmonique généralisée. En 1933, Haar démontra le théorème suivant, qui est le point de départ de toute la théorie : il existe sur un tel groupe une mesure qui est invariante par multiplication à gauche par les éléments du groupe. À partir de ce résultat, le mathématicien soviétique Pontriaguine construisit sa théorie des caractères pour les groupes commutatifs localement compacts, dont un des aspects les plus spectaculaires est sans doute le théorème de dualité. Essayons d'expliquer ce résultat en quelques mots : un caractère d'un groupe topologique G est un homomorphisme continu de G dans le groupe multiplicatif des nombres complexes de module 1 ; il est clair que l'ensemble des caractères forme un groupe commutatif X et on montre que si G est commutatif localement compact, le groupe X peut être muni de manière naturelle d'une structure de groupe topologique localement compact. Le théorème de dualité s'exprime alors par le fait que le groupe G est isomorphe, algébriquement et topologiquement, au groupe des caractères du groupe X.

Issue directement de la théorie des espaces de Banach, la belle théorie des algèbres normées (algèbres de Banach), développée à partir de 1940 par le mathématicien soviétique Gelfand et ses élèves, allait éclairer d'un jour nouveau la dualité de Pontriaguine et permettre d'obtenir d'importants résultats sur la représentation linéaire des groupes localement compacts généraux (et en particulier des groupes de Lie).

JEAN-LUC VERLEY

Bibliographie

D. ALLOUCH & B. CHARLES, *Algèbre générale*, P.U.F., 1984 / N. BOURBAKI, *Éléments d'histoire des mathématiques*, Masson, Paris, 1960 / J. DIEUDONNÉ et al., *Abrégé d'histoire des mathématiques*, Hermann, nouv. éd. 1986 / R. GODEMENT, *Cours d'algèbre*, *ibid.*, 3e éd. 1980 / B. L. VAN DER WAERDEN, *A History of Algebra. From al Khwarizmi to Emmy Noether*, 1985.

ALGÈBRE DE BOOLE
→ BOOLE ALGÈBRE & ANNEAU DE

ALGÈBRE LINÉAIRE & MULTILINÉAIRE
→ LINÉAIRE & MULTILINÉAIRE ALGÈBRE

ALGÈBRE TOPOLOGIQUE
→ TOPOLOGIQUE ALGÈBRE

ALGÈBRES NORMÉES
→ NORMÉES ALGÈBRES

ALGÉBRIQUES NOMBRES
→ NOMBRES (THÉORIE DES) - Nombres algébriques

ANALYSE COMBINATOIRE
→ COMBINATOIRE ANALYSE

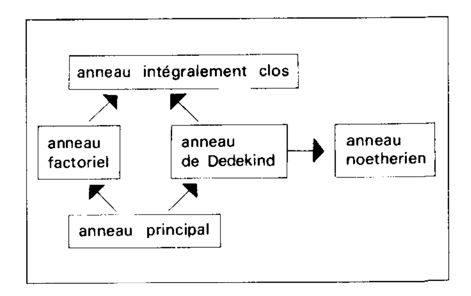
ANALYSE HARMONIQUE
→ HARMONIQUE ANALYSE

ANNEAU DE BOOLE
→ BOOLE ALGÈBRE & ANNEAU DE

ANNEAUX COMMUTATIFS

Dans tout ce qui suit, on se bornera à considérer des anneaux commutatifs unitaires, c'est-à-dire possédant un élément unité pour la multiplication, noté 1. Les définitions sont celles de l'article suivant, ANNEAUX ET ALGÈBRES.

De nombreux cas particuliers d'anneaux commutatifs unitaires ont été étudiés au XIXe siècle, principalement à propos de recherches de théorie des nombres et de géométrie algébrique. Introduits à l'origine pour étudier la divisibilité dans de tels anneaux, les idéaux, cas particuliers de modules, se sont révélés essentiels dans de nombreuses questions. En fait, la classification des différents



types d'anneaux s'effectue suivant la structure de leurs idéaux.

L'arithmétique des anneaux dits *principaux* est analogue à l'arithmétique des nombres entiers ou des polynômes ; plus généralement, on peut étudier de manière satisfaisante l'arithmétique des *anneaux de Dedekind* : ici, les propriétés de divisibilité, déroutantes a priori, s'expriment harmonieusement dans le cadre de la théorie des idéaux. Une autre généralisation possible des anneaux principaux, qui englobe d'ailleurs la précédente, est liée à des conditions de finitude : tout idéal d'un anneau principal est formé des multiples d'un élément ; plus généralement, on peut considérer les anneaux dans lesquels tout idéal est formé des combinaisons linéaires (à coefficients dans l'anneau) d'un nombre fini d'éléments, et ces anneaux, appelés *noethériens*, possèdent une remarquable propriété de stabilité, découverte par Hilbert, à savoir que l'anneau des polynômes sur un anneau noethérien est lui-même noethérien. Pour terminer, mentionnons ici la classe importante des *anneaux locaux*, qui possèdent un unique idéal maximal : cela signifie qu'il existe un idéal propre contenant tous les autres idéaux propres de l'anneau ; ces anneaux jouent un grand rôle dans la théorie des variétés algébriques, différentiables ou analytiques, car les *anneaux de germes de fonctions* sont de ce type. L'étude des anneaux locaux est très liée à des considérations topologiques ; nous renvoyons à ce propos aux articles algèbre TOPOLOGIQUE et théorie des NOMBRES – Nombres *p*-adiques.

Le tableau ci-dessus précise les rapports entre ces différents anneaux, chaque flèche exprimant qu'une propriété en entraîne une autre.



1. Notions fondamentales

Divisibilité

La présence dans un anneau de diviseurs de zéro, c'est-à-dire d'éléments a et b, tous deux non nuls, dont le produit est nul, rend illusoire toute théorie satisfaisante de la divisibilité. Les anneaux commutatifs sans diviseurs de zéro sont appelés des anneaux *intègres* ou *anneaux d'intégrité*. Nous allons, dans ce qui suit, préciser quelques propriétés de la divisibilité dans un tel anneau d'intégrité A. Dans toutes ces questions de divisibilité, seul intervient le fait que l'ensemble A* des éléments non nuls de l'anneau A est muni d'une loi de composition interne $(x, y) \mapsto xy$ (la multiplication) associative, commutative, avec un élément unité ; un ensemble muni d'une loi possédant ces propriétés est appelé un *monoïde*. Nous énoncerons les définitions générales relatives à la divisibilité dans le cadre d'un monoïde A* quelconque, ce qui sera utile dans la troisième partie.

On dit qu'un élément b de A* *divise* un élément a de A*, ou encore que a est *divisible* par b s'il existe un élément c tel que $a = bc$. Il est clair que cette notion de divisibilité généralise la notion usuelle de divisibilité dans le monoïde Z* des entiers relatifs non nuls et possède des propriétés analogues : par exemple, si c divise b et si b divise a, alors c divise a.

Dans toutes les questions de divisibilité, un rôle essentiel est joué par les *unités*, qui sont les éléments inversibles (ou encore, avec la terminologie ci-dessus, les éléments qui divisent l'élément unité 1) ; si A* est le monoïde des éléments non nuls d'un anneau d'intégrité A, ces éléments sont aussi appelés les unités de l'anneau : par

exemple, dans l'anneau **Z** des entiers relatifs, les seules unités sont les nombres $+1$ et -1 et, dans l'anneau des polynômes à coefficients dans un corps K, ce sont les polynômes constants non nuls. Dans tous les cas, on vérifie facilement que les unités forment un groupe multiplicatif ; pour un anneau A, la structure de ce groupe est une importante caractéristique arithmétique de A. Deux éléments a et b, qui diffèrent seulement par un élément inversible, c'est-à-dire tels que $a = ub$, u inversible, possèdent des propriétés de divisibilité très analogues et sont dits *associés*. Pour terminer ces définitions, indiquons qu'un élément a de A* est dit *premier*, ou *irréductible*, s'il n'est pas inversible et si pour toute décomposition $a = bc$; b, c éléments de A*, l'un des deux facteurs b ou c est inversible. Un des problèmes fondamentaux de la divisibilité dans A* est l'étude de la décomposition éventuelle de tout élément comme produit d'éléments premiers.

Corps des fractions d'un anneau d'intégrité

La construction du corps **Q** des nombres rationnels à partir de l'anneau **Z** des entiers relatifs se généralise sans difficulté à un anneau d'intégrité quelconque. Plus précisément, on a le résultat suivant : « Si A est un anneau d'intégrité, il existe un corps K contenant A comme sous-anneau et dont tous les éléments sont de la forme xy^{-1}, x, y éléments de A. De plus, un tel corps K est unique à un isomorphisme laissant A fixe près. »

Pour faire comprendre la démonstration, analysons ce qu'est un nombre rationnel. Un nombre rationnel u est « défini » par une fraction p/q, où p et q sont des entiers relatifs, mais deux fractions p/q et p'/q', distinctes, possédant des numérateurs et des dénominateurs distincts, peuvent définir le *même* nombre rationnel si $pq' = p'q$. De plus, si p/q et p'/q' sont des fractions définissant des nombres rationnels u et v, les fractions $(pq'+p'q)/qq'$ et pp'/qq' définissent les nombres rationnels $u+v$ et uv. La démonstration générale est calquée sur la construction ci-dessus ; donnons-en l'esquisse.

Nous allons d'abord définir la notion de « fraction ». Pour cela, désignons par A* l'ensemble des éléments non nuls de A et considérons l'ensemble A $\times$ A* des couples (x, y), $y \neq 0$; un tel élément (x, y) s'appelle une « fraction » de numérateur x et de dénominateur y. Nous allons maintenant identifier des fractions (x, y) et (x', y') telles que $xy' = x'y$, c'est-à-dire considérer sur l'ensemble A $\times$ A* la relation d'équivalence ainsi définie. L'ensemble des classes d'équivalence forme un ensemble que nous désignerons par K. On vérifie alors facilement que, si on *définit* la somme et le produit de deux « fractions » par les formules :

$$(x, y) + (x', y') = (xy' + x'y, yy')$$
$$(x, y)\,(x', y') = (xx', yy'),$$

on obtient sur K, *par passage au quotient*, deux opérations qui en font un corps ; cela signifie que, si u et u' sont des éléments de K représentés par des « fractions » (x, y) et (x', y'), alors par définition, $u+u'$ et uu' sont les éléments de K représentés par les « fractions » $(x, y) + (x', y')$ et $(x, y)\ (x', y')$ et que $u + u'$ et uu' ainsi définis dont indépendants du choix des « fractions » représentant u et v.

Le plongement de A dans K s'effectue maintenant en identifiant tout élément de A à l'élément de K défini par la « fraction » $(a, 1)$, dont le numérateur est égal à a et le dénominateur à 1. Remarquons que, si on

identifie deux éléments a et b de A à leur image dans K, l'élément de K représenté par la « fraction » (a, b) est bien le quotient (dans K) de a par b.

Le corps K que nous venons de construire s'appelle le *corps des fractions* de l'anneau A.

Idéaux

Rappelons qu'un *idéal* d'un anneau A est un sous-groupe additif qui est stable par multiplication par un élément quelconque de A, qu'il possède certaines propriétés. Nous nous contenterons de montrer comment on peut étendre aux idéaux le langage arithmétique usuel relatif aux nombres entiers.

Les idéaux du type le plus simple sont obtenus ainsi : si a est un élément d'un anneau A, l'ensemble des multiples de a, c'est-à-dire l'ensemble des éléments de la forme xa pour x parcourant A, est un idéal de A, noté (a), et appelé l'idéal *principal* engendré par a. Un tel idéal est propre c'est-à-dire non réduit à 0 et différent de A tout entier si, et seulement si, a est non nul et non inversible. On verra, dans la deuxième partie, que tout idéal de l'anneau Z des entiers relatifs est de ce type. Remarquons au passage que, dans un anneau d'intégrité, deux éléments a et b non nuls engendrent le même idéal principal si et seulement s'ils sont associés, c'est-à-dire si $b = ua$, u inversible dans l'anneau : en effet, si $(a) = (b)$, il existe des éléments u et v tels que $b = ua$, $a = vb$, d'où $vua = a$; si $a \neq 0$, on a donc $vu = 1$ puisqu'il n'y a pas de diviseurs de zéro dans l'anneau et ainsi u est inversible. Plus généralement, si a_1, ..., a_n sont des éléments de A, l'ensemble, noté $(a_1, ..., a_n)$ des éléments de la forme $x_1a_1 + ... + x_na_n$ pour x_1 ..., x_n parcourant A indépendamment l'un et l'autre est un idéal ; la quatrième partie est consacrée à l'étude des anneaux dans lesquels tout idéal est de ce type.

Étant donné deux idéaux $\mathfrak{a}$ et $\mathfrak{b}$, leur intersection $\mathfrak{a} \cap \mathfrak{b}$ est encore un idéal. Généralisons aux idéaux la notion de produit : $\mathfrak{a}$ et $\mathfrak{b}$ étant deux idéaux, l'ensemble des sommes finies $a_1b_1 + ... + a_nb_n$, où les a_i et les b_i sont des éléments de $\mathfrak{a}$ et $\mathfrak{b}$ respectivement, est encore un idéal, appelé *produit* des idéaux $\mathfrak{a}$ et $\mathfrak{b}$ et noté $\mathfrak{ab}$. Le produit ainsi défini est commutatif, associatif et admet un élément unité qui est l'anneau tout entier $A = (1)$ (parfois appelé, pour cette raison, idéal unité). Si A est un anneau d'intégrité, le produit de deux idéaux non nuls (c'est-à-dire différents de $\{0\}$) est non nul et par suite l'ensemble M(A) des idéaux non nuls est un monoïde pour cette loi de composition ; le monoïde M(A) jouera un rôle très important dans la troisième partie. Remarquons que si $\mathfrak{a} = (a)$ et $\mathfrak{b} = (b)$ sont principaux, alors on a $\mathfrak{ab} = (ab)$ et par suite l'application $a \mapsto (a)$ est un homomorphisme du monoïde A* dans le monoïde M(A) (l'image d'un produit est le produit des images, et l'élément unité a pour image l'élément unité).

Deux éléments a et b de A sont dits *congrus modulo un idéal* $\mathfrak{a}$, et on note :

$$a \equiv b \quad (\text{mod. } \mathfrak{a})$$

si la différence $a - b$ appartient à $\mathfrak{a}$; dans le cas où $\mathfrak{a} = (c)$ est principal, on retrouve la notion usuelle de congruence modulo c. Considérons l'ensemble quotient, noté $A/\mathfrak{a}$, de A par cette relation (c'est manifestement une relation d'équivalence). Si $\bar{a}$ et $\bar{b}$ sont les classes de a et b respectivement, on vérifie que $\overline{a+b}$ et $\overline{ab}$ sont indépendants des représentants a et b choisis et que les deux lois de composition ainsi définies font de $A/\mathfrak{a}$

un anneau commutatif unitaire appelé *anneau quotient* de A par l'idéal $\mathfrak{a}$. Dans le cas où A est l'anneau *Z* des entiers relatifs et $\mathfrak{a}$ l'ensemble (n) des multiples d'un entier n, cet anneau n'est autre que l'anneau des classes résiduelles d'entiers modulo n.

Un idéal $\mathfrak{p} \neq$ A est dit *premier* s'il ne contient le produit ab de deux éléments de A que lorsqu'il contient au moins l'un d'entre eux ; dans l'anneau Z des entiers, cette condition caractérise les idéaux principaux (p) engendrés par un nombre premier p. On voit facilement qu'un idéal est premier si, et seulement si, l'anneau quotient est sans diviseurs de zéro ; ainsi, un exemple important d'idéaux premiers est constitué par les idéaux maximaux $\mathfrak{p}$ (idéaux qui ne sont contenus dans aucun autre idéal propre) caractérisés par le fait que A/$\mathfrak{p}$ est un corps. Généralisons maintenant aux idéaux quelconques les propriétés des idéaux principaux de Z engendrés par les puissances des nombres premiers : un idéal $\mathfrak{q}$ est dit *primaire* si $ab \in \mathfrak{q}$ et $a \notin \mathfrak{q}$ entraînent qu'une puissance de b appartient à $\mathfrak{q}$, il résulte des définitions que si $\mathfrak{q}$ est primaire, son *radical*, qui est l'ensemble des éléments dont une puissance appartient à $\mathfrak{q}$, est premier. Nous verrons, dans la quatrième partie, l'importance des idéaux primaires.

Éléments entiers

Soit A un anneau d'intégrité contenu dans un corps K. On dit qu'un élément de K est *entier sur A* s'il est racine d'un polynôme :

$$x^n + a_1x^{n-1} + \ldots + a_n$$

à coefficients dans A et dont le coefficient dominant est égal à 1. Il est clair que tout élément de A est entier sur A puisqu'il est racine du polynôme $x - a$; on peut montrer que l'ensemble des éléments de K qui sont entiers sur A forme un anneau (qui contient donc A) appelé la *fermeture intégrale* de *A dans K*. Un cas particulièrement important est celui où K est le corps des fractions de A (cf. *supra*) ; si les seuls éléments du corps des fractions de A qui sont entiers sur A sont les éléments de A, on dit que l'anneau A est *intégralement clos*. Ces anneaux jouent un rôle essentiel dans de nombreuses questions, en théorie des nombres et en géométrie algébrique notamment.

2. L'arithmétique élémentaire et les anneaux principaux

Un anneau *principal* est un anneau d'intégrité dans lequel tout idéal est principal, c'est-à-dire formé des multiples d'un même élément, appelé *générateur* de l'idéal. L'étude de la divisibilité dans un tel anneau est analogue à la théorie arithmétique élémentaire des nombres entiers, qui en constitue d'ailleurs un cas particulier. L'étude de la divisibilité dans l'anneau K[X] des polynômes à une variable sur un corps K rentre aussi dans ce cadre.

Exemples

a) Montrons que l'*anneau* **Z** *des entiers relatifs est principal*. La démonstration repose sur la propriété suivante de divisibilité dans cet anneau : étant donné deux entiers rationnels a et b, $b > 0$, il existe un couple et un seul d'entiers rationnels q et r tels que :

$$a = bq + r, \qquad 0 \leqslant r < b ;$$

les nombres q et r s'appellent respectivement le quotient et le reste de la division de a par b. Soit donc maintenant $\mathfrak{a}$ un idéal de **Z**. Si $\mathfrak{a} = \{0\}$, on a $\mathfrak{a} = (0)$; sinon $\mathfrak{a}$

contient des éléments strictement positifs puisque avec tout élément a il contient son opposé $-a = (-1)a$. Soit b le plus petit élément strictement positif de $\mathfrak{a}$; montrons que tout élément a de $\mathfrak{a}$ est un multiple de b. En effet, l'existence de la division dans Z permet d'écrire $a = bq + r$, $0 \leqslant r < b$; or le multiple bq de b appartient à $\mathfrak{a}$, donc aussi $r = a - bq$: la définition de b entraîne $r = 0$.

b) Un autre exemple important d'anneau principal est l'anneau K[X] des polynômes à coefficients dans un corps commutatif K. La démonstration repose ici encore sur l'existence dans cet anneau d'une division « euclidienne » : si A et B sont des polynômes, il existe un couple et un seul de polynômes Q et R tels que $A = BQ + R$, le degré de R étant strictement inférieur au degré de B. On montre alors, par une démonstration analogue à ce qui précède, qu'un idéal $\mathfrak{a} \neq (0)$ de K[X] est formé des multiples de tout polynôme B non nul de $\mathfrak{a}$ dont le degré est le plus petit possible (cf. POLYNÔMES).

c) À propos de recherches sur les formes quadratiques, Gauss a utilisé le fait que l'anneau des nombres complexes de la forme $a + bi$, $a, b \in \mathbf{Z}$ (appelés entiers de Gauss), possède une arithmétique comparable à celle des entiers ordinaires. Ce fait s'explique, avec la terminologie ci-dessus, par le fait que cet anneau est principal.

Plus grand commun diviseur et plus petit commun multiple

Dans ce qui suit, nous nous limiterons, pour simplifier les notations, au cas de deux éléments, mais il est clair que tous les résultats s'étendent sans difficulté au cas d'un nombre fini d'éléments.

Soient x, y deux éléments d'un anneau principal A et considérons l'idéal (x, y) constitué par les éléments de la forme $ax + by$, $a, b \in A$. Puisque A est principal, cet idéal est engendré par un élément d, déterminé à cela près qu'on peut le remplacer par ud, où u est un élément inversible quelconque de l'anneau. On appelle *plus grand commun diviseur* (en abrégé P.G.C.D.) de x et y tout élément d tel que $(x, y) = (d)$.

Puisque $d \in (d)$, on voit en particulier qu'il existe $a, b \in A$ tels que :

$$(*) \qquad d = ax + by$$

(ce résultat constitue le *théorème de Bezout*). Justifions la terminologie adoptée en montrant qu'un élément z de A divise simultanément x et y si et seulement s'il divise d : puisque (d) contient x et y, ces nombres sont des multiples de d et, par suite, tout diviseur de d divise x et y ; réciproquement, si z divise x et y, écrivons $x = zx'$, $y = zy'$, et portons dans (*) ; on obtient $d = z(ax' + by')$, ce qui prouve que z divise d. Dans le cas de l'anneau **Z** des entiers rationnels, d est déterminé au signe près puisque les seuls éléments inversibles sont ici $+1$ et -1 ; on peut donc prendre $d > 0$ et on retrouve la notion élémentaire de P.G.C.D. enseignée dans les classes primaires.

Deux éléments x et y de A sont dits *premiers entre eux* s'ils admettent 1 pour P.G.C.D., c'est-à-dire si leurs seuls diviseurs communs sont les unités de l'anneau. D'après le théorème de Bezout indiqué ci-dessus, x et y sont premiers entre eux si et seulement s'il existe des éléments a, $b \in A$ tels que $ax + by = 1$ (la condition suffisante résulte du fait que, si cette relation est satisfaite, tout diviseur commun à x et y divise 1). Ce résultat entraîne facilement le *théorème de Gauss* (ou *lemme d'Euclide*), qui s'énonce : « Soit x et y des

éléments non nuls d'un anneau principal A et d un diviseur du produit xy ; si d et x sont premiers entre eux, alors d divise y. » En effet, puisque d et x sont premiers entre eux, il existe u, $v \in$ A tels que :

$$ud + vx = 1$$

d'où après multiplication par y,

$$y = yud + vxy$$

puisque d divise xy et yud, il divise aussi y.

Soit encore x et y des éléments d'un anneau principal A. Les multiples de x et de y sont les éléments de (x) et (y) respectivement et par suite les multiples communs à x et y sont les éléments de l'idéal $(x) \cap (y)$. Puisque l'anneau A est principal, cet idéal est formé des multiples d'un élément défini à un facteur inversible près : on appelle plus petit commun multiple (en abrégé P.P.C.M.) de x et y tout élément m de A tel que $(x) \cap (y) = (m)$; avec cette définition, tout multiple de x et y est un multiple de m.

Décomposition en facteurs premiers

Pour tout anneau d'intégrité, on a défini sous le titre 1 les éléments premiers. On peut montrer que les anneaux principaux possèdent les deux propriétés fondamentales (F_1) et (F_2) suivantes.

(F_1) *Décomposition en facteurs premiers.* Tout élément x non nul et non inversible est produit d'un nombre fini d'éléments premiers (pas nécessairement distincts).

(F_2) *« Unicité » de la décomposition.* Si

$$x = p_1 p_2 \cdots p_n = p'_1 p'_2 \cdots p'_m$$

sont deux décompositions d'un élément x comme produit d'éléments premiers, alors $m = n$, et, quitte à modifier éventuellement l'ordre des facteurs, les éléments p_i et p'_i sont associés (c'est-à-dire $p'_i = u_i p_i$, u_i inversible), pour $i = 1,2,...,n$.

Remarquons que, puisque deux éléments de A engendrent le même idéal si et seulement s'ils sont associés, les conditions (F_1) et (F_2) expriment que tout idéal de A non nul et différent de A (de la forme (x) puisque A est principal) s'écrit, de *manière unique* à l'ordre près des facteurs, comme un produits d'idéaux premiers :

$$(x) = (p_1)(p_2) \dots (p_n) ;$$

ainsi la situation est plus simple dans le monoïde M(A) que dans le monoïde A* puisqu'il n'y a plus cette fois d'ambiguïté quant au sens à donner à l'expression *unicité de la décomposition*. Dans la pratique, on élimine cette ambiguïté en choisissant une fois pour toutes un ensemble P d'éléments premiers de A tels que pour tout élément premier p' de A il existe un élément premier p de P et un seul qui soit associé à p'. Les propriétés (F_1) et (F_2) s'expriment alors ainsi : tout élément x de A s'écrit, de manière unique à l'ordre près des facteurs, sous la forme :

$$x = up_1 p_2 \cdots p_n$$

où u est une unité de A et où les p_i sont des éléments de P (pas nécessairement distincts). Ainsi, dans le cas de l'anneau **Z** des entiers relatifs, on peut prendre pour P l'ensemble des nombres premiers positifs, et tout entier relatif x s'écrit de manière unique sous la forme :

$$x = \varepsilon p_1 p_2 \cdots p_n, \qquad \varepsilon = \pm 1.$$

Anneaux factoriels

De manière générale, on appelle *anneau factoriel* tout anneau d'intégrité possédant les propriétés (F_1) et (F_2) ; remarquons

d'ailleurs que l'on peut remplacer la condition (F_2) par la condition suivante :

(F_3) Si un élément premier de A divise un produit, il divise au moins un des facteurs de ce produit.

Les anneaux factoriels constituent une classe plus vaste que celle des anneaux principaux ; cette classe possède la remarquable propriété de stabilité suivante : si A est un anneau factoriel, alors l'anneau A[X] des polynômes à coefficient dans A est lui aussi factoriel. On obtient ainsi, par récurrence, que l'anneau $K[X_1,...,X_n]$ des polynômes à n variables sur un corps K est factoriel, alors qu'il n'est pas principal pour $n \geqslant 2$ (en effet, dans l'anneau K[X,Y] des polynômes à deux variables, l'idéal (X,Y), formé des polynômes de la forme $XP(X,Y) + YQ(X,Y)$, n'est pas principal). L'anneau K[[X]] des séries formelles à coefficients dans un corps K est factoriel ; mais, par contre, l'anneau A[[X]] peut ne pas être factoriel même si A est un anneau factoriel (contre-exemple dû à Samuel).

Pour terminer, signalons, en liaison avec la définition donnée plus haut, que *tout anneau factoriel est intégralement clos.* En effet, soit xy^{-1}, x, y éléments de A, un élément du corps des fractions de A ; on peut supposer que x et y n'ont pas de facteurs premiers communs. Si xy^{-1} est entier sur A, il est racine d'un polynôme à coefficients dans A dont le coefficient dominant est égal à 1, soit :

$$x^n y^{-n} + a_1 x^{n-1} y^{-n+1} + ... + a_n = 0,$$

$a_i \in A$; on en déduit que :

$$x^n = -y(a_1 x^{n-1} + ... + a_n y^{n-1}),$$

c'est-à-dire que x^n est un multiple de y. Mais si y n'est pas un élément inversible de A, on obtient une contradiction puisque, d'après (F_3), tout diviseur premier de y doit alors diviser x.

3. Les anneaux de Dedekind et la théorie multiplicative des idéaux

L'extension de l'arithmétique classique aux anneaux d'entiers algébriques s'est longtemps heurtée au fait que ces anneaux ne sont pas factoriels. Par exemple, dans l'anneau $Z[\sqrt{-3}]$ des nombres complexes de la forme $a + ib\sqrt{3}$, a, b entiers relatifs, le nombre 4 admet les deux décompositions :

$$4 = 2 \times 2 = (1 + i\sqrt{3})(1 - i\sqrt{3})$$

en facteurs premiers non associés deux à deux et par suite cet anneau n'est pas factoriel. Dedekind, à partir des travaux de Kummer, mit en évidence que, pour un tel anneau A, la notion importante était celle d'idéal premier et non pas d'élément premier, comme pouvait le faire croire l'étude élémentaire des entiers relatifs. En somme, tout revient ici à remplacer l'étude du monoïde A* des éléments non nuls de A par celle du monoïde M(A) des idéaux non nuls de A ; on trouve l'unicité de la décomposition en facteurs premiers « idéaux ». Chaque élément a non inversible de A* étant identifié à l'idéal principal (a) qu'il engendre peut ainsi s'écrire, de manière unique, comme un produit d'idéaux premiers.

La définition abstraite des anneaux de Dedekind que nous formulons ici a été donnée pour la première fois, en 1927, par la mathématicienne allemande Emmy Noether.

Anneaux de Dedekind

Par définition, on appelle *anneau de Dedekind* tout anneau intégralement clos et noethérien (c'est-à-dire dans lequel tout idéal est engendré par un nombre fini d'éléments, cf. *infra*) dans lequel tout idéal premier non nul est maximal. Cela signifie

que le quotient de A par un idéal premier non nul quelconque est non seulement un anneau d'intégrité mais même un corps.

L'exemple le plus simple d'un tel anneau est l'anneau Z $[\sqrt{d}]$ des nombres de la forme $a + b\sqrt{d}$, a, b entiers relatifs, d entier tel que $d \equiv 2$ ou 3 (mod. 4). Plus généralement, Dedekind a démontré que, si K est une extension finie du corps Q des nombres rationnels (K est appelé un corps de nombres algébriques), alors la fermeture intégrale A de l'anneau Z dans K est un anneau de Dedekind (A est appelé l'anneau des entiers du corps K ; cf. théorie des NOMBRES - Nombres algébriques). En fait, l'exemple précédent, qui est très important en théorie des nombres, est lui-même un cas particulier du résultat algébrique suivant : soit A un anneau de Dedekind, de corps des quotients K, et soit L une extension finie de K (cf. CORPS) ; alors la fermeture intégrale de A dans L est un anneau de Dedekind.

L'intérêt essentiel des anneaux de Dedekind réside dans la structure particulièrement simple, pour un tel anneau, du monoïde M(A) des idéaux non nuls. On a le résultat suivant : un anneau d'intégrité A est un anneau de Dedekind si, et seulement si, tout idéal non nul de A s'écrit de manière unique (à l'ordre près des facteurs) comme produit d'idéaux premiers non nuls. Soit $\mathfrak{a}$ un tel idéal non nul ; écrivant $\mathfrak{p}^a$ si l'idéal premier $\mathfrak{p}$ figure a fois dans la décomposition de $\mathfrak{a}$, on peut donc écrire $\mathfrak{a}$ de manière unique sous la forme :

$$\mathfrak{a} = \mathfrak{p}_1^{a_1}\mathfrak{p}_2^{a_2}\ldots\mathfrak{p}_n^{a_n},$$

les $\mathfrak{p}_i$ étant des idéaux premiers *distincts.* Désignant par P l'ensemble des idéaux premiers non nuls de A, on écrit souvent cette décomposition sous la forme :

$$(1) \qquad \mathfrak{a} = \prod_{\mathfrak{p}\in P} \mathfrak{p}^{v_\mathfrak{p}(a)}$$

en convenant que $v_\mathfrak{p}(\mathfrak{a}) = 0$ quand l'idéal premier $\mathfrak{p}$ ne figure pas dans la décomposition de $\mathfrak{a}$; par définition le produit ci-dessus est alors égal au produit fini correspondant aux idéaux tels que $v_\mathfrak{p}(\mathfrak{a}) > 0$. L'intérêt de cette convention réside dans des formules du type suivant : si $\mathfrak{a}$ et $\mathfrak{b}$ sont deux idéaux, alors on a :

$$\mathfrak{a}\mathfrak{b} = \prod_{\mathfrak{p}\in P} \mathfrak{p}^{v_\mathfrak{p}(a)+v_\mathfrak{p}(b)},$$

c'est-à-dire, avec les notations ci-dessus, $v_\mathfrak{p}(\mathfrak{a}\mathfrak{b}) = v_\mathfrak{p}(\mathfrak{a}) + v_\mathfrak{p}(\mathfrak{b})$.

Remarquons que l'existence et l'unicité de la décomposition de tout idéal de M(A) comme produit d'idéaux premiers permet d'appliquer à M(A) tous les résultats élémentaires relatifs à la divisibilité des entiers ; par exemple si $\mathfrak{a}$ et $\mathfrak{b}$ sont deux idéaux non nuls, leurs diviseurs communs, ou leurs multiples communs, sont les diviseurs, ou les multiples, d'éléments appelés respectivement le P.G.C.D. et le P.P.C.M. de $\mathfrak{a}$ et $\mathfrak{b}$ et qui s'écrivent :

$$\prod_{\mathfrak{p}\in P} \mathfrak{p}^{\inf(v_\mathfrak{p}(a),\, v_\mathfrak{p}(b))}, \quad \prod_{\mathfrak{p}\in P} \mathfrak{p}^{\sup(v_\mathfrak{p}(a),\, v_\mathfrak{p}(b))}$$

respectivement.

Idéaux fractionnaires

Soit A un anneau d'intégrité ; la structure du monoïde M(A) des idéaux non nuls donne des indications sur la structure du monoïde multiplicatif A*. De la même manière, pour étudier la structure du groupe multiplicatif K* du corps des quotients K de A, on est amené à étendre la notion d'idéal.

Un sous-ensemble $\mathfrak{a}$, non réduit à $\{0\}$, de K est appelé un *idéal fractionnaire* si c'est un sous-anneau de K stable par multiplication par les éléments de A et pour lequel il existe un élément $d \neq 0$ de A tel que dx appartienne à A pour tout x

de $\mathfrak{a}$; cette définition revient à dire qu'un idéal fractionnaire est l'ensemble des produits $d^{-1}y$, d non nul fixé, lorsque y parcourt un idéal (au sens usuel) de A. Il est clair qu'un idéal non nul est un idéal fractionnaire ; dans la théorie des idéaux fractionnaires, on appelle souvent *idéaux entiers* les idéaux usuels pour éviter toute confusion.

Il est facile de généraliser aux idéaux fractionnaires les opérations usuelles sur les idéaux : en particulier, si $\mathfrak{a}$ et $\mathfrak{b}$ sont des idéaux fractionnaires, on vérifie que l'ensemble des sommes finies d'éléments du type xy, x dans $\mathfrak{a}$ et y dans $\mathfrak{b}$, est un nouvel idéal fractionnaire appelé le produit de $\mathfrak{a}$ et $\mathfrak{b}$ noté $\mathfrak{ab}$; l'ensemble I(A) des idéaux fractionnaires non nuls est ainsi un monoïde pour cette multiplication. Un idéal fractionnaire $\mathfrak{a}$ est dit *inversible* s'il existe un idéal fractionnaire $\mathfrak{b}$ tel que $\mathfrak{ab} = A$. Avec cette terminologie, on peut maintenant donner une nouvelle définition des anneaux de Dedekind : « Un anneau de Dedekind est un anneau d'intégrité tel que tout idéal fractionnaire non nul est inversible. » Le monoïde I(A) des idéaux fractionnaires non nuls est alors un *groupe* multiplicatif et tout élément de ce groupe peut s'écrire, de manière unique à l'ordre des facteurs près, sous la forme :

$$\mathfrak{a} = \prod_{\mathfrak{p} \in P} \mathfrak{p}^{v_\mathfrak{p}(\mathfrak{a})} \tag{2}$$

où les $v_\mathfrak{p}(\mathfrak{a})$ sont des entiers *relatifs*, nuls sauf pour un nombre fini d'idéaux premiers. Dans cette décomposition, les idéaux entiers sont caractérisés par le fait que $v_\mathfrak{p}(\mathfrak{a}) \geqslant 0$ pour tout idéal premier non nul.

Valuations et idéaux premiers

Soit A un anneau de Dedekind. Pour tout élément $a \neq 0$ de A, l'idéal principal (a) a une décomposition du type (1) : posons, par définition, $v_\mathfrak{p}(a) = v_\mathfrak{p}((a))$ et étendons cette fonction au corps K des fractions de A en posant :

$$v_\mathfrak{p}(xy^{-1}) = v_\mathfrak{p}(x) - v_\mathfrak{p}(y),$$
$$v_\mathfrak{p}(0) = +\infty.$$

On vérifie que, pour $z = xy^{-1}$ dans K, la valeur ainsi définie est indépendante de la décomposition xy^{-1} choisie ; le nombre $v_\mathfrak{p}(z)$ ainsi défini n'est autre que l'exposant de $\mathfrak{p}$ dans la décomposition (2) de l'idéal fractionnaire (z) (formé des éléments az pour a parcourant A). On obtient ainsi une valuation (discrète normée) sur K, c'est-à-dire une fonction v définie sur K et à valeurs dans $\mathbf{Z} \cup \{+\infty\}$ telle que :

(*a*) pour z parcourant K*, $v(z)$ parcourt l'ensemble $\mathbf{Z}$ des entiers relatifs et $v(0) = +\infty$;

(*b*) $v(z_1 z_2) = v(z_1)v(z_2)$;

(*c*) $v(z_1 + z_2) \geqslant \min(v(z_1), v(z_2))$.

Ainsi, à tout idéal premier $\mathfrak{p}$ non nul, on a fait correspondre une valuation, dite *essentielle* ; par définition, une telle valuation prend des valeurs positives ou nulle sur tout élément de A. Réciproquement, on peut montrer que toute valuation (au sens ci-dessus) positive sur A est obtenue à partir d'un idéal premier ; il existe ainsi une correspondance biunivoque entre les valuations (discrètes normées) positives sur A et les idéaux premiers non nuls de l'anneau A. Un exemple est donné par l'anneau $\mathbf{Z}$ des entiers relatifs et l'anneau $\mathbf{Z}_p$ des nombres p-adiques (cf. théorie des NOMBRES - Nombres p-adiques).

4. Anneaux noethériens

Avant Hilbert, les mathématiciens connaissaient fort peu de résultats sur les

anneaux de polynômes à plusieurs variables. À propos de recherches sur la théorie des invariants, Hilbert mit en évidence le fait que tout idéal d'un tel anneau est engendré par un nombre fini d'éléments et montra tout le parti que l'on pouvait tirer de cette propriété ; par là même, il dégageait l'importance des anneaux avec conditions de finitude qui allaient être étudiés systématiquement sous forme générale par E. Noether. Signalons que les conditions de finitude en un sens plus large jouent un rôle absolument essentiel dans toutes les recherches « géométriques » contemporaines en géométrie algébrique ou analytique (au sens moderne, à savoir l'étude des espaces analytiques) et dans de nombreuses questions d'algèbre homologique.

Définitions équivalentes

Un anneau *noethérien* est un anneau commutatif unitaire A qui vérifie une des trois conditions de finitude équivalentes suivantes :

Condition (a), dite de chaîne ascendante : « Toute suite strictement croissante d'idéaux est finie », ou encore : « Si :

$$\mathfrak{a}_1 \subset \mathfrak{a}_2 \subset \ldots \subset \mathfrak{a}_n \subset \ldots$$

est une suite infinie d'idéaux de A encastrés, il existe un entier n tel que $\mathfrak{a}_n = \mathfrak{a}_{n-1} = \ldots$ ».

Condition (b) : « Toute famille non vide d'idéaux a un élément maximal », ce qui signifie que si $(\mathfrak{a}_i)_{i \in \mathrm{I}}$, I non vide fini ou non, est une famille d'idéaux de A, il existe un indice i_0 pour lequel l'idéal $\mathfrak{a}_{i_0}$ n'est contenu strictement dans aucun autre idéal de la famille.

Condition (c) : « Tout idéal est engendré par un nombre fini d'éléments », c'est-à-dire que si $\mathfrak{a}$ est un idéal de A, il existe des éléments $x_1, \ldots, x_n$, en nombre fini, tels que $\mathfrak{a}$ soit l'ensemble des éléments de la forme $a_1x_1 + \ldots + a_nx_n$ lorsque $a_1, \ldots, a_n$ parcourent A indépendamment l'un de l'autre.

Esquissons la démonstration de l'équivalence de ces trois conditions car elle est instructive. Pour montrer que (a) entraîne (b), on raisonne par l'absurde. Si la famille $(\mathfrak{a}_i)_{i \in \mathrm{I}}$ n'admettait pas d'élément maximal, alors pour tout $i \in \mathrm{I}$ on pourrait trouver $j \in \mathrm{I}$ tel que l'idéal $\mathfrak{a}_j$ contienne strictement l'idéal $\mathfrak{a}_i$ et on pourrait construire ainsi, par récurrence à partir d'un idéal $\mathfrak{a}_i$, une suite infinie strictement croissante d'idéaux.

Montrons que (b) entraîne (c). Soit $\mathfrak{a}$ un idéal et considérons la famille des idéaux de type fini (c'est-à-dire engendré par un nombre fini d'éléments) contenus dans $\mathfrak{a}$; cette famille est non vide car elle contient $\{0\}$ et par suite elle admet un élément maximal $\mathfrak{b}$ engendré par des éléments $x_1, \ldots, x_n$. Pour tout x de $\mathfrak{b}$, l'idéal engendré par les éléments $x_1, \ldots, x_n, x$ contient $\mathfrak{b}$, appartient à la famille d'idéaux considérée et par suite est égal à $\mathfrak{b}$ puisque $\mathfrak{b}$ est maximal. Ainsi $\mathfrak{a} = (x_1, \ldots, x_n)$.

Pour terminer, montrons que (c) entraîne (a). Soit $\mathfrak{a}_n$ une suite croissante d'idéaux encastrés. On vérifie que, dans ce cas, la réunion des idéaux $\mathfrak{a}_n$ est encore un idéal ; d'après (c), cet idéal est engendré par un nombre fini d'éléments $x_1, \ldots, x_n$ et, par définition d'une réunion, il existe des entiers $p_1, \ldots, p_n$ tels que $x_1 \in \mathfrak{a}_{p_1}, \ldots, x_n \in \mathfrak{a}_{p_n}$. Il est maintenant clair que, si p est le plus grand des entiers $p_1, \ldots, p_n$, on a $\mathfrak{a}_k = \mathfrak{a}$ pour $k \geqslant p$.

Exemples d'anneaux noethériens

Par définition, les anneaux de Dedekind, et en particulier, bien entendu, les anneaux principaux, sont des anneaux noethériens. Une source importante d'exemples ne rentrant pas dans les précédents est la

remarquable propriété de stabilité suivante, découverte par Hilbert : si A est un anneau noethérien, l'anneau A[X] des polynômes à coefficients dans A est lui aussi noethérien ; par récurrence, ce résultat s'étend à l'anneau $A[X_1, ..., X_n]$ des polynômes à n variables sur un anneau noethérien A. On obtient ainsi que l'anneau des polynômes à n variables à coefficients dans un corps K est noethérien, alors que cet anneau n'est ni principal, ni même de Dedekind, pour $n \geqslant 2$. Signalons que le théorème de Hilbert s'étend à l'anneau $A[X_1, ..., X_n]$ des séries formelles à coefficients dans un anneau noethérien A ; dans le même ordre d'idées, si K est le corps des nombres réels ou des nombres complexes, l'anneau $K\{X_1, ..., X_n\}$ des séries convergentes est noethérien (cf. ANNEAUX ET ALGÈBRES).

Décomposition primaire

Remarquons que, pour un anneau commutatif avec unité, les opérations d'intersection et de produit de deux idéaux jouent des rôles assez semblables. Dans le cas des anneaux principaux par exemple, si p et q sont deux éléments premiers, alors $(p)(q) = (p) \cap (q)$ et par suite la décomposition d'un idéal en idéaux premiers peut s'écrire indifféremment :

$$\begin{aligned}(x) &= (p_1)(p_2)\dots(p_n)\\ &= (p_1)\cap(p_2)\cap\dots\cap(p_n)\,;\end{aligned}$$

cette situation est d'ailleurs la même pour un anneau de Dedekind quelconque. Dans le cas d'un anneau noethérien, le monoïde multiplicatif M(A) n'est guère utilisable, mais on peut donner un théorème de décomposition de tout idéal comme intersection d'idéaux d'un type plus général que les idéaux premiers, les idéaux primaires (cf. *supra*).

Le théorème de décomposition de Lasker-Noether affirme que tout idéal d'un anneau nœthérien s'écrit sous la forme :

$$\mathfrak{a} = \mathfrak{q}_1 \cap \mathfrak{q}_2 \cap \dots \cap \mathfrak{q}_n,$$

où les $\mathfrak{q}_i$ sont des idéaux primaires auxquels on peut imposer les deux conditions suivantes ; aucun des idéaux $\mathfrak{q}_i$ ne contient l'intersection des autres et les radicaux $\mathfrak{p}_i$ des idéaux $\mathfrak{q}_i$ sont des idéaux premiers distincts (cf. *supra*, chap. 1). Il n'y a pas unicité pour une telle décomposition, mais les idéaux premiers $\mathfrak{p}_i$ définis ci-dessus sont déterminés de manière unique et appelés les *idéaux premiers de l'idéal* $\mathfrak{a}$. Parmi ces idéaux premiers, ceux qui sont minimaux, c'est-à-dire qui ne contiennent aucun des autres sont dits *isolés* et ont une grande importance en géométrie algébrique. Cette terminologie est justifiée par le fait que, dans le cas où $\mathfrak{a}$ est un idéal de l'anneau $K[X_1, ..., X_n]$ des polynômes à n variables sur un corps K, ces idéaux isolés correspondent aux composantes irréductibles de l'ensemble des points K^n où s'annulent simultanément tous les polynômes de l'idéal.

JEAN-LUC VERLEY

Bibliographie

N. BOURBAKI, *Éléments de mathématique. Algèbre commutative*, Masson, nouv. éd., 1985 / R. GODEMENT, *Cours d'algèbre*, Hermann, 3e éd. 1980 / S. LANG, *Undergraduate Algebre*, Springer-Verlag, New York, 1987 / A. LEGOFF, *Cours d'algèbre*, Ellipses, 1987 / P. SAMUEL & O. ZARISKI, *Commutative Algebra*, Princeton, vol. II, 1976.

ANNEAUX & ALGÈBRES

Définis par des axiomes qui dégagent les propriétés usuelles des opérations d'addition et de multiplication dans les ensembles de nombres ou les polynômes, les anneaux constituent le cadre général dans lequel on peut appliquer les règles du calcul algébrique élémentaire. Nous donnerons dans cet article les définitions générales et des exemples. Pour une étude plus détaillée des anneaux qui interviennent en théorie des nombres ou en géométrie algébrique, nous renvoyons à l'intérieur de ce texte à d'autres articles.



1. Définitions

Anneaux

Un *anneau* A est un ensemble muni de deux lois de composition internes $(x, y) \mapsto x + y$ et $(x, y) \mapsto xy$, appelées addition et multiplication respectivement, qui possèdent les propriétés suivantes :

(a) $$x + (y + z) = (x + y) + z$$
(associativité de l'addition) ;

(b) $$x + y = y + x$$
(commutativité de l'addition) ;

(c) existence d'un élément, noté 0, tel que, pour tout élément x de A on ait :

$$x + 0 = x$$
(élément neutre pour l'addition) ;

(d) existence, pour tout x de A, d'un élément, noté $-x$, tel que :

$$x + (-x) = 0$$
($-x$ est appelé l'opposé de x) ;

(e) $$x(yz) = (xy)z$$
(associativité de la multiplication) ;

(f) $$x(y + z) = xy + xz$$
$$(y + z)x = yx + zx$$
(double distributivité de la multiplication par rapport à l'addition) ;

(g) bien que cela ne soit pas toujours ainsi dans la littérature, nous supposerons l'existence d'un élément unité pour la multiplication, souvent noté 1, tel que :

$$1x = x1 = x.$$

Les propriétés (a) à (d) expriment que A est un groupe commutatif pour l'addition.

Dans de nombreux exemples, la multiplication est de plus commutative, c'est-à-dire $xy = yx$; un tel anneau est alors dit *commutatif*. Cependant on ne peut pas se limiter à ce cas, car des anneaux importants dans la pratique, les anneaux de matrices par exemple, ne possèdent pas cette propriété ; comme on le verra au début du chapitre, le calcul algébrique dans de tels anneaux réclame quelques précautions. Pour terminer, indiquons qu'un cas particulier très important est constitué par les anneaux dans lesquels tout élément non nul est inversible, c'est-à-dire a un inverse pour la multiplication ; un tel anneau s'appelle un corps (cf. CORPS).

Un sous-ensemble B non vide d'un anneau A est appelé un *sous-anneau*, s'il contient l'unité multiplicative et $x - y$ et xy pour tout couple d'éléments x et y de B ; B est alors un anneau pour les restrictions à B de l'addition et de la multiplication.

Algèbres

Nous introduirons maintenant ici une autre structure qui se rencontre dans de nombreuses questions.

Soit K un corps commutatif. On dira qu'un ensemble E est une K-*algèbre*, ou une *algèbre sur* K, si c'est un espace vectoriel sur le corps K muni d'une application, noté ici multiplicativement :

$$E \times E \to E$$

qui est *bilinéaire*, c'est-à-dire linéaire par rapport à chaque facteur pris séparément :

(a) $(\lambda x + \mu y) . z = \lambda (x . y) + \mu (y . z)$,

(b) $x . (\lambda y + \mu z) = \lambda (x . y) + \mu (x . z)$,

quels que soient les éléments x, y, z de E et les « scalaires » λ et μ appartenant à K. On peut aussi définir une telle structure lorsque K n'est plus un corps, mais seulement un anneau commutatif.

Si la multiplication est associative, on parle d'*algèbre associative* ; cependant certains auteurs oublient de le préciser et incluent l'associativité dans la définition d'une algèbre, mais précisent quand il n'y a pas associativité.

Homomorphismes d'anneaux et algèbres

Soient A et B deux anneaux (ou deux algèbres) et f une application de A dans B. Conformément aux définitions générales des morphismes, on dira que f est un *homomorphisme* d'anneau (ou d'algèbre) si f respecte la structure d'anneau (ou d'algèbre) :

$$f(x + y) = f(x) + f(y),$$
$$f(xy) = f(x) f(y)$$

(et éventuellement, si A et B sont des algèbres, $f(\lambda x) = \lambda f(x)$ pour tout scalaire λ) pour des éléments x et y quelconques de A ; on impose de plus que l'image par f de l'élément unité de A soit l'élément unité de B. Un cas particulier très important est obtenu lorsque f est une application bijective ; l'application inverse est alors aussi un homomorphisme ; on dit que f est un *isomorphisme* et que A et B sont des anneaux ou des algèbres *isomorphes*. Du point de vue de la théorie des anneaux, il n'y a pas lieu de distinguer entre eux des anneaux isomorphes.

2. Exemples d'anneaux et algèbres

On rencontrera des anneaux et des algèbres dans de nombreux articles du présent ouvrage ; nous nous contenterons donc ici de choisir quelques exemples, de manière un peu artificielle, dans des domaines variés des mathématiques pour montrer la richesse de ces structures.

Les ensembles de nombres sont des exemples très simples d'anneaux pour les opérations usuelles d'addition et de multiplication : l'ensemble **Z** des entiers relatifs est un anneau commutatif unitaire et les ensembles **Q**, **R**, **C**, des nombres rationnels, réels et complexes respectivement, sont des corps. Si A est un anneau commutatif, l'ensemble $A[X_1, ..., X_n]$ des polynômes à n variables à coefficients dans A est un anneau commutatif ; si A = K est un corps, alors l'anneau des polynômes à coefficient dans K est une algèbre sur K.

Un exemple fondamental d'algèbre non commutative est constitué par l'algèbre $\mathcal{L}(E)$ des endomorphismes d'un espace vectoriel E ; si E est de dimension finie n, alors cette algèbre est isomorphe à l'algèbre des matrices carrées d'ordre n, à n lignes et n colonnes (cf. algèbre LINÉAIRE ET MULTILINÉAIRE).

Comme exemple d'algèbre non associative, citons les algèbres de Lie (cf. GROUPES - Groupes de Lie).

Anneaux de Boole

L'exemple suivant montre le caractère un peu insolite que peuvent présenter certains anneaux. L'ensemble $\mathscr{P}(E)$ des parties d'un ensemble donné E est un anneau pour les opérations d'« addition » et de « multiplication » qui à deux sous-ensembles X et Y de E font correspondre les sous-ensembles :

$$(X \cap Y') \cup (X' \cap Y) \text{ et } X \cap Y$$

respectivement, en désignant par X' et Y' les complémentaires de X et Y dans E ; l'élément nul est ici l'ensemble vide et l'élément unité est l'ensemble E tout entier. Remarquons que le « produit » de X par lui-même est égal à X car on a $X \cap X = X$.

Revenons aux notations usuelles en désignant les éléments d'un anneau par des lettres minuscules. Généralisant la situation précédente, on considère des anneaux, appelés *anneaux de Boole*, qui possèdent la propriété que le carré de tout élément est égal à cet élément : $x^2 = xx = x$. Il en résulte que, pour tout élément x, on a $x + x = 0$; en effet, écrivant que le produit de $x + x$ par lui-même est égal à $x + x$, on obtient :

$$\begin{aligned} x + x &= (x + x)(x + x) \\ &= xx + xx + xx + xx = x + x + x + x, \end{aligned}$$

d'où la conclusion. Ces anneaux sont importants en logique symbolique (algèbre des propositions) et dans la théorie des circuits électroniques (algèbre des circuits).

Anneaux et algèbres de fonctions

Les fonctions réelles d'une variable réelle définies dans un intervalle $[a, b]$ de la droite réelle constituent une algèbre en convenant que la somme et le produit de deux fonctions ou le produit d'une fonction par un nombre réel λ sont les fonctions dont les valeurs en chaque point sont respectivement la somme et le produit des valeurs en ce point ou le produit par λ de la valeur de la fonction en ce point. Si on analyse les propriétés qui ont permis de munir l'ensemble précédent d'une structure d'algèbre, on constate que, de manière générale, on peut munir d'une structure d'anneau ou d'algèbre l'ensemble des applications d'un ensemble quelconque E dans un anneau ou une algèbre respectivement, les valeurs en un point x de E des fonctions somme, produit et éventuellement produit par un scalaire étant données par :

$$(f + g)(x) = f(x) + g(x), \quad (fg)(x) = f(x)g(x), \\ (\lambda . f)(x) = \lambda . f(x).$$

Le procédé précédent permet, bien entendu, de définir des structures d'anneaux ou d'algèbres sur de nombreux ensembles de fonctions contenus dans l'ensemble, considéré ci-dessus, de *toutes* les fonctions définies dans un ensemble et à valeurs dans un anneau ou une algèbre. Ainsi, les fonctions continues ou différentiables à valeurs réelles définies dans un ouvert du plan constituent des algèbres sur le corps des nombres réels. Il est clair qu'il est possible de multiplier à volonté les exemples de ce type.

Lorsqu'on s'intéresse à l'étude locale des fonctions au voisinage d'un point, on est conduit à introduire des anneaux et algèbres d'un type différent du précédent. Nous prendrons pour exemple l'algèbre des *germes de fonctions analytiques à l'origine* O du plan complexe. Considérons les couples (U, f) d'un voisinage ouvert de O dans le plan complexe et d'une fonction f définie et analytique dans U. Nous dirons que deux tels couples (U, f) et (V, g) définissent le même germe à l'origine si f et g coïncident

sur un voisinage ouvert W de O contenu dans $U \cap V$; il est clair que cette relation est une relation d'équivalence ; par définition, le germe d'une fonction analytique définie dans un voisinage de O est sa classe d'équivalence pour cette relation. Montrons que cet ensemble des germes peut être muni d'une structure d'algèbre sur le corps des nombres réels ou des nombres complexes. Soient A et B deux germes et λ un nombre réel ou complexe. Si (U, f) et (V, g) définissent les germes A et B respectivement, nous appellerons germe somme, germe produit et germe produit par le scalaire λ, noté $A + B$, AB et λA respectivement, les germes à l'origine des couples $(U \cap V, f_1 + g_1)$, $(U \cap V, f_1 g_1)$, et $(U \cap V, \lambda f_1)$, où f_1 et g_1 désignent les restrictions au voisinage $U \cap V$ des fonctions f et g ; on vérifie alors facilement que les germes $A + B$, AB et λA sont indépendants des représentants (U, f) et (V, g) choisis et que l'ensemble des germes est ainsi muni d'une structure d'algèbre. De manière générale, les anneaux de germes de fonctions différentiables ou analytiques jouent un rôle absolument essentiel dans la théorie des variétés différentiables ou analytiques.

Anneaux de séries

A étant un anneau commutatif, on peut définir de manière purement formelle et algébrique des séries à coefficients dans A ; dans le cas où A est le corps des nombres complexes ou des nombres réels, nous ferons jouer un rôle particulier à celles de ces séries, dites convergentes, qui possèdent un rayon de convergence non nul.

On appelle *série formelle* (à une variable) à coefficients dans un anneau commutatif A une suite infinie d'éléments de A : $(a_0, a_1, ..., a_n, ...)$; une telle série formelle est souvent notée :

$$\sum_{p \geqslant 0} a_p X^p,$$

notation qu'il faut considérer pour l'instant comme un pur symbole. Définissons la somme et le produit de deux telles séries formelles ; on pose, par définition,

$$\sum_{p \geqslant 0} a_p X^p + \sum_{p \geqslant 0} b_p X^p = \sum_{p \geqslant 0} (a_p + b_p) X^p,$$

$$\Big(\sum_{p \geqslant 0} a_p X^p\Big)\Big(\sum_{p \geqslant 0} b_p X^p\Big) = \sum_{p \geqslant 0} c_p X^p,$$

où c_p est la somme finie :

$$a_0 b_p + a_1 b_{p-1} + ... + a_k b_{p-k} + ... + a_p b_0.$$

Il est facile maintenant de vérifier, en utilisant les règles du calcul algébrique dans les anneaux (cf. chap. 3), que l'ensemble A [[X]] de ces séries formelles est muni d'une structure d'anneau ; si A = K est un corps, cet anneau est une algèbre quand on définit la multiplication scalaire par la formule :

$$\lambda\Big(\sum a_p X^p\Big) = \sum (\lambda a_p) X^p.$$

Par récurrence, on peut définir l'anneau des séries formelles à n variables à coefficients dans A ; par définition, cet anneau, noté A $[[X_1, ..., X_n]]$, est égal à l'anneau des séries formelles (à une variable) à coefficients dans l'anneau A $[[X_1, ..., X_{n-1}]]$ des séries formelles à $(n-1)$ variables. Toute série formelle à n variables est définie par la donnée, pour tout système de n entiers $p_1, ..., p_n$ positifs ou nuls, d'un élément $a_{p_1 \cdots p_n}$ de l'anneau A et s'écrit symboliquement sous la forme :

$$(*) \qquad \sum_{p \geqslant 0} \cdots \sum_{p \geqslant 0} a_{p_1 \ldots p_n} X^{p_1} \ldots X^{p_n}.$$

Limitons-nous maintenant au cas où A est le corps des nombres réels ou des nombres complexes. La série (*) est dite *convergente* si elle a un rayon de convergence non nul, c'est-à-dire s'il existe un nombre réel strictement positif R tel que la famille de nombres positifs :

$$|a_{p_1 \dots p_n}| \mathbf{R}^{p_1 + \dots + p_n}$$

soit sommable (cela signifie qu'il existe un nombre M qui majore toute somme finie de tels nombres). On montre que si deux séries sont convergentes, alors les séries formelles somme et produit sont aussi des séries convergentes ; ainsi les séries convergentes à coefficients dans le corps des réels ou des nombres complexes forment des anneaux, qui sont d'ailleurs aussi des algèbres sur **R** ou **C**, notés $\mathbf{R}\{X_1, \dots, X_n\}$ et $\mathbf{C}\{X_1, \dots, X_n\}$ respectivement. L'étude de ces anneaux constitue la partie locale de la théorie des fonctions analytiques de plusieurs variables ; ainsi, montrons, pour $n = 1$ par exemple, que l'anneau $\mathbf{C}\{X\}$ des séries convergentes à coefficients complexes est isomorphe à l'anneau des germes de fonctions analytiques à l'origine introduit ci-dessus. En effet, toute fonction analytique dans un voisinage de l'origine est développable en série entière convergente et deux telles fonctions définissent le même germe si, et seulement si, elles sont somme d'une même série entière dans un voisinage de l'origine ; par ailleurs, la valeur, pour z complexe assez voisin de 0, de la somme des séries « somme » et « produit » est égale respectivement à la somme et au produit des sommes des séries considérées (cf. FONCTIONS ANALYTIQUES).

Algèbres de dimension finie

Soit A une algèbre sur un corps K dont l'espace vectoriel sous-jacent soit de dimension finie n et choisissons une base $e_1, \dots, e_n$ de cet espace. On appelle *table de multiplication* de A la donnée des produits :

$$e_i e_j = \sum_{k=1}^{n} a_{ijk} e_k$$

(les n^3 éléments a_{ijk} de K ainsi définis sont appelés les constantes de structure de l'algèbre A) ; connaissant la table, on peut calculer le produit de deux éléments quelconques par bilinéarité. On représente souvent la table par un schéma à double entrée. Par exemple, la table de multiplication du corps des nombres complexes considéré comme une algèbre de dimension 2 sur le corps des nombres réels est la suivante :

	1	i
1	1	i
i	i	-1

Réciproquement, soit E un espace vectoriel muni d'une base $e_1, \dots, e_n$. Si on se donne, pour tout couple i, j d'entiers compris entre 1 et n, des éléments de l'espace E, notés e_i, e_j, on peut prolonger cette loi par bilinéarité à l'espace E tout entier. L'espace E est alors une algèbre associative admettant cette loi pour multiplication si, et seulement si, on a $(e_i e_j) e_k = e_i (e_j e_k)$; remarquons que l'algèbre ainsi construite est commutative si, et seulement si, $e_i e_j = e_j e_i$. Ce qui précède montre l'utilité des tables pour définir des algèbres. Nous allons donner un exemple célèbre de cette situation.

Soit K un corps commutatif ; désignons par e, i, j, k la base canonique de l'espace vectoriel K^4 et choisissons deux éléments p et q de K. On appelle *algèbre de quaternions* sur K l'algèbre obtenue en considérant sur K^4 la table de multiplication notée **H** :

$$e^2 = e,\ ei = ie = i,\ ej = je = j,\ ek = ke = k,$$
$$i^2 = pe,\ j^2 = qe,\ k^2 = -pqe,$$
$$ij = -ji = k,\ jk = -kj = -qi,$$
$$ki = -ik = -pj.$$

Un cas particulier très important s'obtient en prenant pour K le corps des nombres réels et en choisissant $p = q = -1$; on obtient ainsi les *quaternions* proprement dits, introduits par Hamilton. Pour ces quaternions, on peut développer une théorie analogue à celle des nombres complexes : si $x = ae + bi + cj + dk$ est un tel quaternion, on appelle conjugué de x le quaternion $\bar{x} = ae - bi - cj - dk$; les règles de calcul montrent alors que :

$$x\bar{x} = a^2 + b^2 + c^2 + d^2,$$

et par suite tout quaternion non nul x a un inverse :

$$x^{-1} = (a^2 + b^2 + c^2 + d^2)^{-1}\bar{x},$$

tel que $xx^{-1} = x^{-1}x = e$. On traduit cette propriété en disant que les quaternions forment un *corps non commutatif*; cet exemple des quaternions constitue une situation très privilégiée, car on peut montrer que c'est le seul corps non commutatif de dimension finie sur le corps des nombres réels.

Pour terminer, remarquons que les matrices de la forme :

$$\begin{pmatrix} a + ib & c + id \\ -c + id & a - ib \end{pmatrix}$$

où a, b, c, d sont des nombres réels quelconques forment une algèbre et que l'application qui, au quaternion $ae + bi + cj + dk$, fait correspondre la matrice ci-dessus est un isomorphisme car les opérations usuelles sur les matrices correspondent ici aux opérations correspondantes sur les quaternions ; ainsi l'algèbre des quaternions est isomorphe à une algèbre de matrices. La recherche d'algèbres de matrices isomorphes à une algèbre donnée est le problème fondamental de la représentation linéaire des algèbres ; la situation précédente constitue historiquement le premier exemple d'une telle représentation.

3. Propriétés des anneaux et algèbres

Calcul algébrique dans les anneaux

Les règles du calcul algébrique usuel s'appliquent dans les anneaux moyennant quelques précautions dans le cas non commutatif ; par exemple, si $x_1, ..., x_m, y_1, ..., y_n$ sont des éléments d'un anneau A, le produit

$$(x_1 + ... + x_m)\,(y_1 + ... + y_n)$$

est égal à la somme des mn produits x_iy_j. Mentionnons une importante notation qui montre qu'on peut faire « opérer » l'anneau **Z** des entiers relatifs sur un anneau A quelconque. Si n est un entier relatif et x un élément de A, on désigne par nx la somme d'une suite de n termes égaux à x si $n > 0$, l'élément 0 si $n = 0$ et l'opposé de la somme de $n' = -n$ termes égaux à $-x$ si $n < 0$; il est clair que cette notation possède les propriétés habituelles :

$$n\,(mx) = (nm)x,\ (n + m)x = nx + mx,$$
$$n\,(x + y) = nx + ny,\ (nx)\,(my) = (nm)\,(xy),$$

pour m, n dans **Z** et x, y dans A.

L'exemple des anneaux de Boole montre qu'il peut exister dans certains anneaux des entiers $n > 0$ tels que $n1 = 0$; on appelle *caractéristique* d'un tel anneau le plus petit entier $n > 0$ pour lequel $n1 = 0$ et on dit qu'un anneau est de caractéristique nulle si $n1 \neq 0$ pour tout $n > 0$. Ainsi tout anneau de Boole est de caractéristique 2, alors que l'anneau des entiers relatifs est de caractéristique nulle ; de manière générale, tout anneau de caractéristique nulle contient une infinité d'élé-

ments (si n et m sont deux entiers relatifs distincts les éléments $n1$ et $m1$ sont distincts car $(n-m)\,1 \neq 0$) et par suite tout anneau ne contenant qu'un nombre fini d'éléments est de caractéristique non nulle. Pour terminer avec les notations, indiquons qu'on désigne par x^n, n entier > 0, le produit d'une suite de n termes égaux à x ; il est clair que deux telles puissances de x vérifient :

$$x^n x^m = x^m x^n = x^{n+m}.$$

Remarquons que, si x et y sont deux éléments quelconques d'un anneau A, on a :

$$(x+y)(x-y) = x^2 - y^2 + xy - yx\,;$$

si x et y *commutent* : $xy = yx$, alors on retrouve la formule classique :

$$(x+y)(x-y) = x^2 - y^2.$$

Cette situation se généralise aux *identités remarquables,* qui sont valables si les éléments qui y figurent commutent. Par exemple, on a :

$$x^n - y^n = (x-y)(x^{n-1} + x^{n-2}y + \dots + xy^{n-2} + y^{n-1}),$$

$$(x+y)^n = x^n + C_n^1 x^{n-1}y + \dots + C_n^p x^{n-p}y^p + \dots + y^n$$

(formule du binôme) si x et y commutent.

Puisque pour n premier tous les coefficients C_n^1, C_n^2, C_n^{n-1} sont des entiers divisibles par n, il résulte de la formule du binôme que si x et y sont deux éléments qui commutent dans un anneau de caractéristique n premier, on a $(x+y)^n = x^n + y^n$; d'autre part, sous les mêmes hypothèses, on a $(xy)^n = x^n y^n$. Ainsi dans un anneau commutatif de caractéristique n premier l'application $x \mapsto x^n$ est un homomorphisme d'anneau.

Dans un anneau quelconque, il n'est pas toujours possible de « simplifier par a » une égalité du type $ax = ay$. Ainsi, dans un anneau de Boole unitaire, on a toujours $x^2 - x = x(x-1) = 0$ et, par suite, le produit de deux éléments non nuls peut être nul ; de même, dans l'anneau (de caractéristique nulle) des fonctions à valeurs réelles définies sur l'ensemble réunion de deux ensembles X et Y sans point commun, le produit de deux fonctions l'une nulle sur X et non nulle sur Y et l'autre nulle sur Y et non nulle sur X est nul (cf. chap. 2). De manière générale, on dit qu'un élément $x \neq 0$ d'un anneau A est un *diviseur de zéro* (à gauche) s'il existe un élément $y \neq 0$ tel que $xy = 0$. Un cas particulier de cette situation est constitué par les éléments non nuls dont une puissance est nulle (ainsi, dans l'anneau des entiers modulo 4, cf. *infra*, le carré de la classe du nombre 2 est la classe nulle) ; un tel élément non nul dont une puissance est nulle est appelé un élément *nilpotent.* Les anneaux commutatifs sans diviseurs de zéro sont dits *intègres* (on dit aussi qu'un tel anneau est un *anneau d'intégrité*) ; on peut alors « simplifier » par un élément a non nul puisque $ax = ay$ est équivalent à $a(x-y) = 0$, qui entraîne $x - y = 0$.

Idéaux

Soient A et B deux anneaux (ou deux algèbres) et f un homomorphisme d'anneau (ou d'algèbre) de A dans B. L'ensemble N des éléments de A dont l'image par f est l'élément nul de B est appelé le *noyau* de f ; c'est un sous-groupe additif (ou une sous-algèbre) de A qui possède la propriété supplémentaire suivante : « Pour tout élément x de A et tout élément y de N, les éléments xy et yx appartiennent encore à N. » De manière plus générale, on appelle *idéal à gauche* d'un anneau (ou d'une algèbre) A tout sous-groupe additif (ou sous-algèbre) $\mathfrak{A}$ tel

que si x et y sont des éléments quelconques de A et $\mathfrak{A}$ respectivement, xy soit un élément de $\mathfrak{A}$. On définirait de même les idéaux à droite caractérisés par le fait que yx appartient à $\mathfrak{A}$ pour y dans $\mathfrak{A}$ et x dans A. Un ensemble qui, comme le noyau d'un homomorphisme, est à la fois un idéal à droite et un idéal à gauche est appelé un *idéal bilatère* ; bien entendu, si l'anneau A est commutatif, tous ces types d'idéaux coïncident. Tout anneau contient au moins deux idéaux bilatères particulièrement simples, l'*idéal nul* contenant seulement l'élément 0 et l'*idéal unité* constitué par l'anneau tout entier ; un idéal distinct de ces deux idéaux est dit *propre*. On vérifie facilement que si un anneau est un corps, il n'a pas d'idéaux propres. Donnons deux exemples simples d'idéaux propres : dans l'anneau des entiers relatifs, les multiples d'un nombre n forment un idéal, noté (n) et appelé l'idéal principal engendré par n (cet idéal n'est pas nul si $n \neq 0$ et est différent de **Z** si $n \neq \pm 1$) et on peut montrer que tout idéal est de ce type (cf. ANNEAUX COMMUTATIFS) ; de même, dans un anneau de fonctions, l'ensemble des fonctions qui s'annulent en un point est un idéal.

Indiquons maintenant un procédé souvent utilisé pour construire des idéaux ; idéal signifiera ici indifféremment idéal à gauche, à droite ou bilatère, sauf précision complémentaire. On voit facilement que l'intersection d'une famille quelconque d'idéaux, finie ou non, est encore un idéal. On en déduit que si M est une partie quelconque de A, l'intersection des idéaux de A qui contiennent M (il existe au moins un tel idéal, à savoir A tout entier) est un idéal $\mathfrak{A}$, qui est le plus petit idéal contenant M ; on dit alors que M est un système de générateurs de $\mathfrak{A}$, ou encore que $\mathfrak{A}$ est engendré par M. Par exemple, l'idéal à gauche engendré par l'ensemble contenant un seul élément a est l'ensemble des éléments de la forme xa lorsque x parcourt A, c'est-à-dire l'ensemble des « multiples à gauche » de a ; de manière générale, l'idéal à gauche engendré par une partie M de A, finie ou non, est identique à l'ensemble des sommes finies $x_1a_1 + ... + x_na_n$ où (a_i) est une famille *finie* quelconque d'éléments de M et les x_i des éléments quelconques de A.

Un idéal $\mathfrak{A} \neq$ A d'un anneau A est dit *maximal* s'il n'est contenu dans aucun autre idéal propre de A. Déterminons à titre d'exemple les idéaux maximaux de l'anneau **Z** des entiers relatifs. Admettons ici (cf. ANNEAUX COMMUTATIFS) que tout idéal de **Z** est égal à l'ensemble (n) des multiples d'un élément n ; on voit que l'idéal (n) contient l'idéal (m) si et seulement si m lui-même est un multiple de n ; ainsi l'idéal (p) est maximal si et seulement s'il n'existe pas d'entier $n \neq p$ divisant p et tel que $(n) \neq$ A, c'est-à-dire $n \neq \pm 1$; ainsi, dans l'anneau **Z**, les idéaux maximaux sont formés des multiples des nombres premiers. L'intérêt de la notion d'idéal maximal résulte surtout du théorème suivant, dû à Krull, et dont la démonstration utilise l'axiome du choix : « Dans un anneau, tout idéal à gauche différent de l'anneau tout entier est contenu dans au moins un idéal à gauche maximal. » Nous renvoyons à l'article algèbres NORMÉES pour voir un bel exemple de l'utilité de la notion d'idéal maximal.

L'exemple des germes de fonctions, ou des anneaux de séries, nous montre l'importance des anneaux possédant un unique idéal maximal, qui est alors maximum, c'est-à-dire qui contient tous les idéaux de l'anneau distincts de l'anneau lui-même. Dans le cas de l'anneau des germes de fonctions analytiques à l'origine, les germes dont un représentant

(U, f) s'annule pour $z = 0$ (il en est alors de même de *tous* les représentants d'un tel germe) forment manifestement un idéal. Cet idéal contient tout idéal propre : en effet si g est une fonction analytique dans un voisinage de l'origine qui ne s'annule pas pour $z = 0$, il existe un voisinage U de O dans lequel $g(z) \neq 0$ et, par suite, dans lequel $h(z) = 1/g(z)$ est analytique ; ainsi le germe A, défini par g, a un inverse B dans l'anneau (c'est le germe défini par h) ; tout germe C est alors un multiple de A car $C = (CB)A$ et le seul idéal contenant A est donc l'anneau des germes tout entier. On démontre également que tout anneau de séries possède un unique idéal maximal. Les anneaux de ce type, qui peuvent aussi être caractérisés par le fait que l'ensemble des éléments non inversibles est un idéal, sont appelés des *anneaux locaux*.

Anneau quotient

Les idéaux bilatères d'un anneau (ou d'une algèbre) A jouent un rôle fondamental dans l'étude des relations d'équivalence sur A compatibles avec sa structure d'anneau (ou d'algèbre). De manière précise, toute relation d'équivalence telle qu'on puisse munir l'ensemble quotient d'une structure d'anneau (ou d'algèbre) pour laquelle l'application canonique (qui à un élément fait correspondre sa classe) soit un homomorphisme s'obtient de la façon suivante : il existe un idéal bilatère $\mathfrak{U}$ tel que deux éléments x et y soient équivalents si et seulement si leur différence $x - y$ appartient à l'idéal $\mathfrak{U}$. Si X et Y sont deux classes, on définit leur somme, leur produit, et éventuellement leur produit par un scalaire λ, en choisissant des représentants x et y de X et Y ; on vérifie alors (et ici intervient le fait que $\mathfrak{U}$ est un idéal) que les classes $X + Y$, XY et λX des éléments $x + y$, xy, et, éventuellement λx sont indépendantes des représentants x et y choisis et que l'ensemble quotient est un anneau (ou une algèbre) noté $A/\mathfrak{U}$ pour les opérations ainsi définies. A s'appelle *l'anneau quotient de* A *par l'idéal* $\mathfrak{U}$. Il est clair que si A est commutatif, alors l'anneau quotient par un idéal est encore commutatif. Avec ces notions, un idéal $\mathfrak{U}$ d'un anneau commutatif est maximal si et seulement si l'anneau quotient $A/\mathfrak{U}$ est un corps.

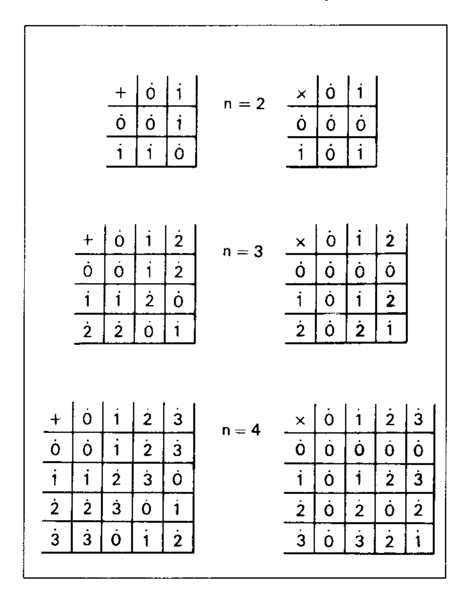
Anneau des entiers relatifs modulo n

Nous allons maintenant indiquer un exemple fondamental d'anneau quotient qui montrera que le calcul des congruences dans l'anneau **Z** des entiers relatifs rentre dans la théorie des anneaux.

Soit n un entier positif et considérons la relation d'équivalence définie par l'idéal $(n) = n\mathbf{Z}$ des multiples de n ; deux entiers x et y sont équivalents pour cette relation d'équivalence si, et seulement si, leur différence est un multiple de n, c'est-à-dire avec la terminologie classique en arithmétique, si x et y sont *congrus modulo n* (cette relation est notée $x \equiv y$, mod. n) ; la classe d'un entier x s'appelle la *classe résiduelle de x modulo n.* D'après les propriétés du quotient d'un anneau commutatif unitaire par un idéal, l'ensemble $\mathbf{Z}/n\mathbf{Z}$ des classes résiduelles modulo n est un anneau commutatif unitaire ; il en résulte en particulier qu'on peut appliquer aux congruences les règles usuelles du calcul algébrique.

Dans l'anneau $\mathbf{Z}/n\mathbf{Z}$, les classes des nombres $0, 1, ..., n - 1$ sont distinctes car la différence de deux tels nombres est inférieure à n en valeur absolue et par suite ne peut être un multiple de n ; réciproquement, tout nombre entier est égal à un de ces nombres à un multiple de n près. Cela

montre que $\mathbf{Z}/n\mathbf{Z}$ est un anneau fini contenant exactement n éléments qui sont les classes $0, 1, ..., n-1$ des nombres $0, 1, ..., n-1$. Le tableau ci-joint donne



$n = 2$

+	$\dot{0}$	$\dot{1}$
$\dot{0}$	$\dot{0}$	$\dot{1}$
$\dot{1}$	$\dot{1}$	$\dot{0}$

×	$\dot{0}$	$\dot{1}$
$\dot{0}$	$\dot{0}$	$\dot{0}$
$\dot{1}$	$\dot{0}$	$\dot{1}$

$n = 3$

+	$\dot{0}$	$\dot{1}$	$\dot{2}$
$\dot{0}$	$\dot{0}$	$\dot{1}$	$\dot{2}$
$\dot{1}$	$\dot{1}$	$\dot{2}$	$\dot{0}$
$\dot{2}$	$\dot{2}$	$\dot{0}$	$\dot{1}$

×	$\dot{0}$	$\dot{1}$	$\dot{2}$
$\dot{0}$	$\dot{0}$	$\dot{0}$	$\dot{0}$
$\dot{1}$	$\dot{0}$	$\dot{1}$	$\dot{2}$
$\dot{2}$	$\dot{0}$	$\dot{2}$	$\dot{1}$

$n = 4$

+	$\dot{0}$	$\dot{1}$	$\dot{2}$	$\dot{3}$
$\dot{0}$	$\dot{0}$	$\dot{1}$	$\dot{2}$	$\dot{3}$
$\dot{1}$	$\dot{1}$	$\dot{2}$	$\dot{3}$	$\dot{0}$
$\dot{2}$	$\dot{2}$	$\dot{3}$	$\dot{0}$	$\dot{1}$
$\dot{3}$	$\dot{3}$	$\dot{0}$	$\dot{1}$	$\dot{2}$

×	$\dot{0}$	$\dot{1}$	$\dot{2}$	$\dot{3}$
$\dot{0}$	$\dot{0}$	$\dot{0}$	$\dot{0}$	$\dot{0}$
$\dot{1}$	$\dot{0}$	$\dot{1}$	$\dot{2}$	$\dot{3}$
$\dot{2}$	$\dot{0}$	$\dot{2}$	$\dot{0}$	$\dot{2}$
$\dot{3}$	$\dot{0}$	$\dot{3}$	$\dot{2}$	$\dot{1}$

les tables d'addition et de multiplication dans les anneaux $\mathbf{Z}/2\mathbf{Z}$, $\mathbf{Z}/3\mathbf{Z}$ et $\mathbf{Z}/4\mathbf{Z}$. On y remarque que $\mathbf{Z}/2\mathbf{Z}$ est un anneau de Boole, car $(\dot{0})^2 = \dot{0}$ et $(\dot{1})^2 = \dot{1}$; réciproquement, on peut montrer que tout anneau de Boole sans diviseur de zéro est isomorphe à l'anneau $\mathbf{Z}/2\mathbf{Z}$. Dans l'anneau $\mathbf{Z}/4\mathbf{Z}$, l'élément 2 est nilpotent car son carré est nul. L'anneau $\mathbf{Z}/3\mathbf{Z}$ est un corps, car tout élément non nul a un inverse dans l'anneau. Montrons plus généralement que l'anneau $\mathbf{Z}/p\mathbf{Z}$ est un corps si, et seulement si, p est un nombre premier. En effet, si p est un nombre entier positif non premier, p est le produit de deux nombres entiers q_1 et q_2 positifs et strictement inférieurs à p ; on a donc $\dot{0} = p = q_1 q_2$ et $\mathbf{Z}/p\mathbf{Z}$ n'est pas un corps car il contient des diviseurs de zéro. Réciproquement, si p est premier, tout nombre $q > 0$ plus petit que p est premier avec p et par suite il existe des entiers relatifs u et v tels que $up + vq = 1$ (théorème de Bezout) ; passant aux classes d'équivalence, on a $\dot{u}\dot{p} + \dot{v}\dot{q} = \dot{v}\dot{q} = \dot{1}$ et par suite $\dot{q}$ est inversible, d'inverse $\dot{v}$; ainsi tout élément non nul de $\mathbf{Z}/(p)$ est inversible et cet anneau est un corps souvent noté F_p (cf. CORPS).

JEAN-LUC VERLEY

Bibliographie

N. BOURBAKI, *Éléments de mathématiques*, livre II, *Algèbre*, chap. I à III, Association Collaborateurs Bourbaki, Paris, 1982 / R. GODEMENT, *Cours d'algèbre*, Hermann, 3[e] éd. 1980 / N. JACOBSON, *Lectures in Abstract Algebra*, 3 vol., Van Nostrand, Princeton, 1977 / S. LANG, *Undergraduate Algebre*, Springer-Verlag, New York, 1987 / C. MUTAFIAN, *Le Défi algébrique* t. II, Vuibert, 1976.

APPLICATION AFFINE
→ AFFINE APPLICATION

APPLICATIONS PROJECTIVES
→ PROJECTIVES APPLICATIONS

APPROXIMATIONS DES FONCTIONS
→ FONCTIONS REPRÉSENTATION & APPROXIMATION DES

APPROXIMATIONS DIOPHANTIENNES → DIOPHANTIENNES APPROXIMATIONS

ASYMPTOTIQUES CALCULS

Il est difficile de définir avec précision ce que l'on appelle méthodes *asymptotiques* en analyse mathématique. Ainsi, lors de l'étude d'une suite ou d'une fonction dont la nature est compliquée, certaines questions ne nécessitent que des renseignements d'ordre qualitatif tels que $f(x) \to 0$ ou $f(x) \to +\infty$ pour $x \to +\infty$. D'autres exigent un contrôle quantitatif très précis, défini par des inégalités explicites. Les comportements asymptotiques relèvent d'une préoccupation intermédiaire : dans de nombreux problèmes, on remplace la quantité étudiée par une autre plus simple sans que, « à la limite », le résultat soit modifié. Par exemple, la relation

$$f(x) \sim \frac{k}{x^2}$$

suffit à établir la convergence à l'infini de l'intégrale de f. Les exemples qui suivent montreront la nature de ces préoccupations.

Du point de vue strictement technique, les méthodes asymptotiques sont extrêmement variées et, en dehors de quelques résultats relativement généraux, chaque cas particulier exerce l'ingéniosité de celui qui l'étudie. Nous nous limiterons, dans les chapitres 2 et 3, à l'exposé de quelques-unes de ces méthodes.

1. Comparaison de la croissance des fonctions

L'étude de la manière dont des quantités tendent vers l'infini ou tendent vers zéro a constitué, à la naissance du calcul infinitésimal, au XVIIe siècle, la théorie des « infiniment grands » et de leurs inverses, les « infiniment petits », et a fait l'objet de polémiques passionnées, souvent paramathématiques. En effet, l'absence d'une conception précise de la notion de limite (problème qui ne fut pas abordé sous forme rigoureuse) rendait mystérieuse la nature de ces quantités qui, tout en étant comparables entre elles, n'étaient pas comparables aux nombres ou aux fonctions. De plus, cela représentait l'intrusion de l'infini dans les calculs et, bien que cet infini fût potentiel et non actuel (en ce sens que l'infini n'était pas considéré en lui-même comme un être mathématique soumis à des règles opératoires précises), cela suscitait des problèmes philosophiques aux mathématiciens, échaudés par les nombreux « paradoxes de l'infini » qui n'étaient pas encore clairement analysés. La recherche des « vraies valeurs » des formes indéterminées, rencontrées dans le calcul des dérivées par exemple, allait cependant montrer l'importance de ces notions et les justifier, tout au moins empiriquement, aux yeux des mathématiciens. Ces questions ont été systématiquement élucidées par Paul Du Bois-Reymond, qui, dans une série d'articles de 1870-1871, a posé les fondements du calcul des infiniment grands (*Infinitär-calcul*) en mettant en évidence l'impor-

tance de la notion d'échelle de comparaison ; il a étudié également en détail l'intégration et la dérivation des relations de comparaison. Tous ces résultats ont trouvé leur forme rigoureuse et définitive dans les travaux de Hardy.

Relations de comparaison

Dans de nombreuses questions d'analyse, on est conduit à étudier l'« ordre de grandeur » d'une fonction $f(n)$ d'une variable entière positive pour n grand, ou d'une fonction $f(x)$ d'une variable continue x lorsque x est grand, ou voisin d'une valeur a. Formulons le problème de manière précise, en nous limitant au cas d'une variable continue pour fixer les idées.

Soit f une fonction, à valeurs réelles ou complexes, définie pour $a - h < x < a$ ou $a < x < a + h$ (on dira alors que f est définie dans un voisinage à gauche ou à droite de a) ou pour x assez grand (on dira que f est définie au voisinage de l'infini). Par exemple, l'application :

$$x \mapsto \frac{1}{x - a},$$

transforme les voisinages à droite de a en les voisinages de l'infini et, par suite, on peut, en considérant la fonction :

$$f\left(\frac{1}{x - a}\right),$$

ramener l'étude de f au voisinage de a (à droite) à l'étude d'une autre fonction au voisinage de l'infini ; nous nous limiterons à ce cas dans ce chapitre. Il est clair, d'autre part, que les notions ci-dessous s'étendent de manière évidente au cas où la variable ne prend que des valeurs entières. Nous nous proposons ici de montrer qu'il est possible de « comparer » les fonctions définies au voisinage de l'infini lorsque la variable tend vers l'infini.

Si f et g sont deux fonctions définies pour x assez grand et à valeurs non nulles, on dit que f et g sont asymptotiquement *équivalentes* pour x tendant vers l'infini, et on note :

$$f(x) \sim g(x),$$

au voisinage de l'infini si le rapport $f(x)/g(x)$ tend vers 1 pour x tendant vers l'infini. Remarquons que cela n'entraîne même pas que la différence $f(x) - g(x)$ soit bornée, comme le montre l'exemple des fonctions asymptotiquement équivalentes $x^2 + x$ et x^2 au voisinage de l'infini. La recherche de fonctions équivalentes à des fonctions données et plus simples ou plus maniables est un problème essentiel (cf. *Partie principale*, in chap. 2) et qui peut être fort difficile ; ainsi, si $\pi(x)$ désigne le nombre de nombres premiers $\leqslant x$, la « loi de répartition des nombres premiers », conjecturée par Gauss, qui exprime que :

$$\pi(x) \sim \frac{x}{\ln x}, \quad x \to \infty,$$

est un résultat profond de la théorie des nombres, qui n'a été démontré qu'en 1896 par Hadamard et de La Vallée-Poussin.

Nous allons maintenant préciser comment on peut comparer les croissances des fonctions pour $x \to \infty$. On dit que f est asymptotiquement *négligeable* devant g, ou que *g croît plus vite que f* pour x tendant vers l'infini, si pour tout $\varepsilon > 0$, on a $|f(x)| \leqslant \varepsilon |g(x)|$ pour x assez grand ; on écrit alors :

$$f(x) = o(g(x)) \quad x \to \infty$$

(notation de Landau).

Il est clair que la définition ci-dessus signifie que le rapport $f(x)/g(x)$, s'il est

défini pour x assez grand, tend vers 0. Remarquons que le symbole de Landau doit être considéré seulement comme une écriture commode mais ne représente pas une fonction déterminée croissant moins vite que g et ne doit donc pas être manipulé comme les nombres ou les fonctions ; ainsi écrire que $o(g) + o(g) = o(g)$ exprime que la somme de deux fonctions négligeables devant g est négligeable devant g, mais cela n'aurait aucun sens d'en conclure que $o(g) = 0$! Pour une formulation mathématique satisfaisante, on pourrait convenir que $o(g)$ désigne l'*ensemble* des fonctions négligeables devant g, qui est un espace vectoriel ; la notation $f = o(g)$ est alors un abus de langage pour $f \in o(g)$.

La notation précédente permet de traduire maintenant facilement les résultats sur la « croissance comparée » des fonctions puissance, logarithme et exponentielle ; ainsi, pour $a < b$,

$$x^a = o(x^b), \quad x \to \infty ;$$

pour a quelconque et $b > 0$, on a :

$$(\ln x)^a = o(x^b), \quad x \to \infty,$$

c'est-à-dire les fonctions puissance croissent plus vite que les fonctions logarithme. Enfin, pour a quelconque, $c > 0$ et $d > 0$,

$$x^a = o(e^{cx^d})$$

au voisinage de l'infini, c'est-à-dire les fonctions exponentielles croissent plus vite que les fonctions puissance.

Indiquons, pour terminer, une dernière notation fort utile. Dans de nombreux cas, on a besoin d'une relation de comparaison plus faible que la précédente, qui puisse s'appliquer à des fonctions dont la croissance n'est pas très régulière. S'il existe une constante M telle que $|f(x)| \leqslant M|g(x)|$ pour x assez grand, on dit que f est dominée par g pour x tendant vers l'infini, et on écrit :

$$f(x) = O(g(x)), x \to \infty \text{ (notation de Landau)} ;$$

bien entendu, si f est négligeable devant g, elle est aussi dominée par g. Si g est la fonction constante 1, l'égalité $f(x) = O(1)$ exprime simplement que f est bornée pour x assez grand ; par exemple :

$$\sin x + \frac{1}{x} = O(1), \quad x \to \infty.$$

Le symbole O permet de préciser certains ordres de croissance ; ainsi écrire que :

$$f(x) = O\left(\frac{1}{x^2}\right), \quad x \to \infty,$$

est plus précis que :

$$f(x) = o\left(\frac{1}{x}\right), \quad x \to \infty.$$

Les Anglo-Saxons utilisent les notations de Hardy :

$$f \ll g \text{ ou } f(x) \ll g(x)$$

pour f est négligeable devant g, ou :

$$f \preccurlyeq g \text{ ou } f(x) \preccurlyeq g(x)$$

pour f est dominée par g.

Fonctions de comparaison et échelles de comparaison

Les fonctions les plus simples auxquelles on est tenté de comparer (au sens du chapitre 1) une fonction donnée, pour x tendant vers l'infini, sont les fonctions :

$$x^a \ (a \text{ réel}), \ (\ln x)^b \ (b \text{ réel} \neq 0),$$
$$e^{cx^d} \ (c \neq 0 \text{ et } d > 0)$$

et les produits d'un nombre fini de telles fonctions. Remarquons au passage que, à l'exception de la fonction 1 (obtenue pour $a = 0$), chacune de ces fonctions tend vers 0 ou vers l'infini lorsque x tend vers

l'infini ; de plus, d'après les propriétés de croissance comparée de ces fonctions, si f et g sont de ce type, on a une et une seule des trois relations :

$$(1) \quad \begin{cases} f(x) = o(g(x)), & f(x) = g(x), \\ g(x) = o(f(x)) & x \to \infty ; \end{cases}$$

cette propriété signifie que deux fonctions distinctes f et g du type considéré ont des croissances que l'on peut comparer et ne font pas « double emploi », en ce sens qu'il n'existe pas de constante c telle que $f(x) \sim cg(x), x \to \infty$. De manière générale, une famille E de fonctions définies et positives pour x assez grand peut être utilisée de manière profitable pour étudier la croissance d'une fonction quelconque si, quelles que soient f et g dans E, on a une et une seule des relations (1) ; on dit alors que la famille E constitue une *échelle de comparaison* pour x tendant vers l'infini. Bien entendu, toute partie d'une échelle de comparaison en est aussi une ; dans la pratique, on se limite à des *échelles logarithmico-exponentielles* constituées de fonctions du type considéré ci-dessus ou de composées (au sens de la composition des fonctions) de telles fonctions, par exemple :

$$L_2 x = \ln \ln x, \quad e_2(x) = \exp(\exp x).$$

On utilise souvent des échelles de comparaison dénombrables ; l'exemple le plus simple est :

$$(2) \quad \ldots, x^n, x^{n-1}, \ldots, x, 1, x^{-1}, \ldots, x^{-n}, \ldots, \quad x \to \infty,$$

qui est *décroissante*, en ce sens que chaque terme est négligeable, pour x tendant vers l'infini, devant tous les termes précédents. Indiquons, pour terminer, que le choix d'une échelle de comparaison dépend essentiellement du type de fonction que l'on veut étudier et il serait illusoire d'espérer trouver une échelle dénombrable telle que toute fonction s'« intercale » exactement dans l'échelle ; en effet, on peut montrer que, pour toute échelle dénombrable, il existe une fonction qui croît plus vite que toute fonction de l'échelle (Du Bois-Reymond) et une fonction qui croît moins vite que toute fonction de l'échelle (Hadamard).

Tout ce qui précède sur les relations de comparaison au voisinage de l'infini se définit de même pour les fonctions définies dans un voisinage (à droite par exemple) d'un point a : ainsi l'échelle de comparaison qui correspond à (2) est dans ce cas :

$$(3) \quad \ldots, (x-a)^{-n}, \ldots, (x-a)^{-1}, 1, x-a, \ldots, (x-a)^n, \ldots$$

2. Développements asymptotiques

Dans ce chapitre, on supposera choisie une échelle de comparaison E (au voisinage d'un point a ou au voisinage de l'infini).

Partie principale

L'idée la plus simple pour étudier le comportement d'une fonction donnée f (au voisinage de a ou de l'infini) est de chercher si, à une constante près, elle est équivalente à une fonction de l'échelle E choisie. S'il existe une telle fonction g de E et une constante $c \neq 0$ telles que $f \sim cg$, c et g sont déterminées de manière unique et on dit que cg est la *partie principale* de f (par rapport à l'échelle E) ; remarquons que cela équivaut à dire que :

$$f(x) = cg(x) + o(g(x)),$$

ou encore que $f(x)/g(x)$ tend vers une limite finie $c \neq 0$. Dans le cas où l'échelle choisie est (2) ou (3), on retrouve la notion usuelle de partie principale ; ainsi, $1/\sin x$ a

pour partie principale $1/x$ pour $x \to 0$, $e^x - e^a$ a pour partie principale $e^a(x-a)$ pour $x \to a$,

$$\frac{2x^2+1}{x-1}$$

a pour partie principale $2x$ pour $x \to \infty$. Remarquons que la partie principale n'existe pas nécessairement ; en effet, toutes les fonctions d'une échelle logarithmico-exponentielle sont positives pour x assez grand et par suite une fonction « oscillante » (comme $x \sin x$, qui s'annule dans tout voisinage de l'infini) n'est comparable à aucune fonction de ce type, pour $x \to \infty$. Il se peut aussi que l'échelle choisie ne soit pas assez « riche » et que f croisse plus vite ou moins vite que toute fonction de l'échelle, ou encore tombe dans un « trou » de l'échelle : ainsi la fonction $x \ln x$ n'a pas de partie principale par rapport à l'échelle (2), car elle croît plus vite que x et moins vite que x^2.

Développements asymptotiques au sens de Poincaré

Si f a une partie principale c_1g_1 par rapport à une échelle E, on peut chercher à préciser un peu plus le comportement de f en étudiant la différence $f - c_1g_1$; si cette fonction a une partie principale c_2g_2, on a alors :

$$f(x) = c_1g_1(x) + c_2g_2(x) + \mathrm{o}(g_2(x)),\quad g_2(x) = \mathrm{o}(g_1(x)).$$

De manière générale, on appelle *développement asymptotique* (au sens de Henri Poincaré) *d'ordre k d'une fonction f par rapport à une échelle de comparaison* E une somme *finie* (nécessairement déterminée de manière unique si elle existe) :

$$g(x) = c_1g_1(x) + c_2g_2(x) + ... + c_ig_i(x) + ... c_kg_k(x),\quad g_i \in \mathrm{E} \text{ et } g_{i+1}(x) = \mathrm{o}(g_i(x))$$

telle que la différence $f - g$ soit négligeable, les g_i formant une suite « décroissante » de fonctions de E, en ce sens que, pour chaque i, la fonction g_{i+1} est négligeable devant g_i.

L'exemple le plus simple de cette situation est la théorie classique des *développements limités* au voisinage d'un point a : ce n'est autre que la recherche du développement asymptotique d'une fonction par rapport à l'échelle (3). Le résultat classique le plus important est ici la *formule de Taylor*, qui affirme que toute fonction k fois continûment dérivable au voisinage de a admet le développement limité d'ordre k :

$$f(x) = f(a) + f'(a)(x-a) + ... + \frac{f^{(k)}(a)}{k!}(x-a)^k + \mathrm{o}((x-a)^k),\quad x \to a.$$

On obtient ainsi, pour les fonctions usuelles de l'analyse, des développements limités d'ordre arbitrairement grand ; par exemple, au voisinage de 0 :

$$(1+x)^s = 1 + sx + ... + \frac{s(s-1)...(s-k+1)}{k!}x^k + \mathrm{o}(x^k),$$

$$e^x = 1 + x + ... + \frac{x^k}{k!} + \mathrm{o}(x^k).$$

On trouvera de nombreux autres exemples dans l'article SÉRIES & PRODUITS INFINIS.

Les résultats précédents permettent déjà d'étudier un grand nombre de formes indéterminées.

3. Cas des intégrales

Il s'agit d'étudier le comportement asymptotique des *restes* des intégrales convergentes :

$$x \mapsto \int_x^\infty f(t)\,dt$$

ou d'évaluer des intégrales divergentes :

$$x \mapsto \int_a^x f(t)\,dt.$$

Cette étude s'effectue en deux étapes. On se ramène au cas où f appartient à une échelle classique de comparaison, grâce au théorème d'intégration des relations de comparaison :

Théorème. Si f et g sont positives et équivalentes au voisinage de $+\infty$, alors :
- dans le cas convergent :

$$\int_x^\infty f(t)\,dt \sim \int_x^\infty g(t)\,dt,$$

- dans le cas divergent :

$$\int_0^x f(t)\,dt \sim \int_0^x g(t)\,dt.$$

Les énoncés sont analogues pour les relations $f = o(g)$ et $f = O(g)$. En revanche, on ne peut pas toujours *dériver* les relations de comparaison ; par exemple :

$$x\left(1 + x\sin\frac{1}{x}\right) \sim x,$$

mais :

$$1 + 2x\sin\frac{1}{x} - \cos\frac{1}{x}$$

n'est pas équivalent à 1.

Pour f appartenant à une échelle classique, si on ne dispose pas d'une primitive explicite, on effectue des intégrations par parties successives. Par exemple, le comportement asymptotique du logarithme intégral :

$$\mathrm{li}(x) = \int_2^x \frac{dt}{\ln t},$$

étudié par Euler et Gauss, est donné par :

$$\int_2^x \frac{dt}{\ln t} = \frac{x}{\ln x} + \frac{1\,!x}{(\ln x)^2} + \dots + \frac{(k-1)!x}{(\ln x)^k} + o\left(\frac{x}{(\ln x)^k}\right).$$

De même, la fonction d'erreur, introduite par Gauss en calcul des probabilités,

$$\mathrm{Er\,f}(x) = \frac{2}{\sqrt{\pi}}\int_2^x e^{-t^2}\,dt$$

tend vers 1 si $x \to +\infty$ et son développement asymptotique se déduit de la relation :

$$\int_x^{+\infty} e^{-t^2}dt = e^{-x^2}\left(\frac{1}{2x} - \frac{1}{4x^3} + \dots + (-1)^k\frac{1.3\dots(2k-1)}{2^{k+1}x^{2k+1}} + o\left(\frac{1}{x^{2k+1}}\right)\right).$$

Il est important de ne pas confondre les développements asymptotiques avec les séries ; dans de nombreux cas, on n'est capable de déterminer explicitement qu'un petit nombre de termes et d'obtenir cependant ainsi de très précieux renseignements sur les fonctions considérées. De toute façon, un développement asymptotique est essentiellement une somme *finie* et, même si on peut obtenir un nombre arbitrairement grand de termes, cela n'entraîne nullement que la série correspondante converge, comme le montre l'exemple suivant, étudié par Laplace. Considérons la fonction :

$$f(t) = \int_1^t \frac{e^x}{x}\,dx$$

(à une constante près, c'est la fonction « exponentielle-intégrale ») ; par intégration successive par parties, on obtient facilement, pour tout k,

$$e^{-t}f(t) = \frac{1}{t} + \frac{1}{t^2} + \dots + \frac{(k-1)!}{t^k} + o\left(\frac{1}{t^k}\right), \quad t \to \infty;$$

on constate facilement que, pour tout t, la série de terme général $(k-1)!/t^k$ est divergente.

Dans l'exemple du logarithme intégral $\mathrm{li}(x)$ donné ci-dessus, on remarquera même que *tous* les termes du développement tendent vers $+\infty$ avec x !

À travers les exemples précédents, on voit que le rôle de l'intégration par parties est de transformer l'intégrale à étudier en une intégrale négligeable devant la précédente. On peut expliquer le schéma de calcul de la manière suivante. Soit l'intégrale :

$$\int_1^x f(t)g(t)\,dt,$$

où f varie lentement devant g (par exemple $f(t) = 1/t$, ou $f(t) = \ln t$, et $g(t) = e^t$, $g(t) = \exp(-t^2)$, ou $g(t) = e^{it}$). En première approximation, on assimile f à une constante, ce qui conduit à l'estimation :

$$\int_1^x f(t)g(t)\,dt \# f(x)\mathrm{G}(x),$$

où :

$$\mathrm{G}(x) = \int_1^x g(t)\,dt.$$

La non-constance de f se traduit par l'apparition d'un terme correctif portant sur la dérivée de f, ce qui constitue la formule d'intégration par parties :

$$\int_1^x f(t)g(t)\,dt = \Big[f(t)\mathrm{G}(t)\Big]_1^x - \int_1^x f'(t)\mathrm{G}(t)\,dt.$$

La faible variation de f se traduit par le fait que l'intégrale du second membre est négligeable devant l'intégrale initiale. Dans les applications, on rencontre très souvent le cas de l'amortissement constant $g(t) = e^{-kt}$ ou de la phase tournante uniforme $g(t) = e^{i\omega t}$. Ce double aspect amortissement et phase tournante se retrouve dans la méthode de Laplace et dans la méthode de la phase stationnaire (cf. chap. 5).

4. Cas des séries

Généralités

Le cas du développement asymptotique des sommes partielles des séries est analogue à celui des intégrales. Il s'agit d'étudier le comportement asymptotique des restes de séries convergentes :

$$\sum_{p=n+1}^{\infty} f(p)$$

et des sommes partielles de séries divergentes :

$$\sum_{p=0}^{n} f(p).$$

Ici encore, on se ramène au cas où f appartient à une échelle classique, grâce au théorème de sommation des relations de comparaison :

Théorème. Si f et g sont positives et équivalentes au voisinage de $+\infty$, alors :

– dans le cas convergent :

$$\sum_{p=n+1}^{\infty} f(p) \sim \sum_{p=n+1}^{\infty} g(p),$$

– dans le cas divergent :

$$\sum_{p=0}^{n} f(p) \sim \sum_{p=0}^{n} g(p),$$

Enfin, lorsque f appartient à une échelle classique, on compare les sommes précédentes à des intégrales. Cette comparaison est facile lorsque f varie « lentement » ; plus précisément, on a :

Théorème de Hardy. Soit f une fonction à valeurs complexes de classe $\mathcal{C}^1$, pour $x \geqslant 0$, telle que sa dérivée f' soit intégrable au voisinage de $+\infty$. Alors, la suite :

$$\sum_{p=0}^{n} f(p) - \int_0^n f(t)\,dt$$

est convergente.

Par exemple, la suite :

$$\sum_{p=1}^{n} \frac{1}{p} - \ln n$$

admet une limite, traditionnellement notée γ et appelée constante d'Euler. Ce nombre est de nature encore très mystérieuse et on ne sait même pas s'il est rationnel ou irrationnel. Une valeur approchée à 20 décimales est :

0,57721 56649 01532 86060.

Formule sommatoire d'Euler-Maclaurin

Si l'on désire un développement asymptotique à une précision plus grande, on fait appel à une technique beaucoup plus élaborée, la formule sommatoire d'Euler-Maclaurin. On suppose ici que f est suffisamment dérivable et que, pour tout entier k, la dérivée k-ième $f^{(k)}$ est négligeable devant $f^{(k-1)}$.

On se propose d'évaluer des sommes du type :

$$\sum_{n=p+1}^{q} f(n)$$

par comparaison avec l'intégrale :

$$\int_p^q f(t)\, dt.$$

Plus précisément écrivons :

$$\sum_{n=p+1}^{q} f(n) - \int_p^q f(t)\, dt = \sum_{n=p+1}^{q} \left[\int_{n-1}^{n} (f(n) - f(t))\, dt\right];$$

on se ramène d'abord à l'intervalle [0, 1] pour chacune des intégrales de la somme de droite par les changements de variable $u + n - 1 = t$. On a, pour chacune de ces intégrales :

$$\int_{n-1}^{n} [f(n) - f(t)]dt = \int_0^1 [f(n) - f(u+n-1)]\, du$$
$$= [f(n-1) - f(n)]P_1(0) + \int_0^1 f'(u+n-1)P_1(u)\, du$$

après intégration par parties, en prenant une primitive P_1 du polynôme constant $P_0 = 1$. De même, en prenant une primitive P_2 de P_1, on a, par intégration par parties :

$$\int_0^1 f'(u+n-1)P_1(u)\, du = \Big[f'(u+n-1)P_2(u)\Big]_0^1 - \int_0^1 f''(u+n-1)P_2(u)\, du.$$

On poursuit ce processus jusqu'au rang r, c'est-à-dire jusqu'à un reste portant sur $f^{(r+1)}$, et on choisit les primitives successives P_k de $P_0 = 1$ de telle sorte que les termes tout intégrés disparaissent deux à deux dans la sommation lorsque n varie de $p + 1$ à q. Il suffit pour cela d'imposer au polynôme P_k de vérifier les relations : $P'_k = P_{k-1}$ pour $k \geqslant 1$ et $P_k(1) = P_k(0)$ pour $k \geqslant 2$, ou, ce qui revient au même, pour tout $k \geqslant 1$,

$$P'_k = P_{k-1} \quad \text{et} \quad \int_0^1 P_k(u)\, du = 0.$$

On démontre qu'il existe une suite (P_k) et une seule satisfaisant à ces conditions. Plus précisément, pour tout entier naturel k,

$$P_k = \frac{1}{k!} B_k,$$

où B_k *est le* k-ième polynôme de Bernoulli, considéré par Jacques Bernoulli comme solution de l'équation aux différences : $P(x+1) - P(x) = kx^{k-1}$. Ces polynômes peuvent être introduits par la série génératrice formelle :

$$\frac{t\,e^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} B_n(x)\frac{t^n}{n!}.$$

En fait, dans ce qui précède, seuls interviennent les *nombres* de Bernoulli $\beta_k = B_k(0)$. En particulier, $\beta_0 = 1$, $\beta_1 = -1/2$, $\beta_{2p+1} = 0$ si $p \geqslant 1$, $\beta_2 = 1/6$, $\beta_4 = -1/30$, $\beta_6 = 1/42$, $\beta_8 = -1/30$, $\beta_{10} = 5/66$.

Pour expliciter les calculs précédents, il convient de distinguer deux cas, suivant que :

f est intégrable au voisinage de $+\infty$; il s'agit alors d'évaluer le reste :

$$\sum_{n=p+1}^{\infty} f(n).$$

f n'est pas intégrable ; il s'agit alors d'évaluer la somme partielle :

$$\sum_{n=0}^{p} f(n).$$

Il convient de noter que la formule sommatoire d'Euler-Maclaurin est de type asymptotique, c'est-à-dire que les restes ne tendent pas nécessairement vers 0. Mais, lorsque ce reste tend vers 0, la formule sommatoire fournit des développements en série. Ainsi, appliquant cette formule à la fonction $t \mapsto e^{zt}$, on montre que, pour $|z| < 2\pi$,

$$\frac{z}{e^z - 1} = \sum_{n=0}^{\infty} \frac{\beta_n}{n!} z^n.$$

On en déduit les développements en séries entières des fonctions trigonométriques :

$$\operatorname{cotg} z = \frac{1}{z} = \sum_{n=1}^{\infty} (-1)^n 2^{2n} \frac{\beta_{2n}}{(2n)!} z^{2n-1},$$

$$\operatorname{tg} z = \sum_{n=1}^{\infty} (-1)^{n-1} 2^{2n} (2^{2n} - 1) \frac{\beta_{2n}}{(2n)!} z^{2n-1},$$

$$\frac{1}{\sin z} = \frac{1}{z} + \sum_{n=1}^{\infty} (-1)^{n-1} 2(2^{2n-1} - 1) \frac{\beta_{2n}}{(2n)!} z^{2n-1}.$$

La formule d'Euler-Maclaurin s'applique à l'évaluation de sommes portant sur une fonction f dont les dérivées décroissent de plus en plus. Les cas les plus fréquents en mathématiques appliquées sont ceux où f présente un amortissement constant, une phase tournante constante, et la situation mixte. Par exemple, supposons que l'on veuille évaluer :

$$\sum_{n=p+1}^{\infty} \frac{e^{-nz}}{n^\alpha}$$

avec $z = x + iy$, $x \geqslant 0$ $(z \neq 2ik\pi)$ et $\alpha \geqslant 0$. On peut penser à comparer e^{-nz}/n^α à l'intégrale :

$$\int_{n-1}^{n} \frac{e^{-tz}}{t^\alpha} dt.$$

Or, en première approximation, $1/t^\alpha$ varie peu sur $[n-1, n]$ tandis que e^{-tz} présente de manière mixte un amortissement et une phase tournante, ce qui conduit à l'estimation :

$$(1) \qquad \int_{n-1}^{n} \frac{e^{-tz}}{t^\alpha} dt \; \# \; \frac{e^z - 1}{z} \times \frac{e^{-nz}}{n^\alpha}.$$

Autrement dit, le terme e^{-nz}/n^α est multiplié par un facteur constant dû à l'amortissement et à la phase. L'estimation précédente permet de conjecturer :

$$\sum_{n=p+1}^{\infty} \frac{e^{-nz}}{n^\alpha} \sim \frac{z}{e^z - 1} \int_p^\infty \frac{e^{-tz}}{t^\alpha} dt\,;$$

pour établir rigoureusement cette relation, il faut évidemment remplacer l'estimation (1) par une intégration par parties comme indiqué *supra*.

5. Cas des fonctions définies par des intégrales

Nous dégagerons ici trois méthodes importantes pour étudier le comportement asymptotique d'intégrales dépendant d'un paramètre lorsque ce paramètre tend vers l'infini.

La méthode de Laplace

Considérons une fonction :

$$I(t) = \int_a^b g(x)\, e^{th(x)}\, dx,$$

définie par une intégrale ; (a, b) est ici un intervalle quelconque, borné ou pas. Pour simplifier, nous supposerons que la fonction h admet un seul maximum. L'idée essentielle ici est que, sous cette hypothèse, c'est la partie de l'intégrale située au voisinage de ce maximum qui est prédominante pour t grand ; par suite, si on remplace la fonction $g(x)\, e^{th(x)}$ par sa partie principale au voisinage de ce point, il est plausible que l'on obtienne, par intégration, la partie principale de $I(t)$ pour t tendant vers l'infini. Nous nous limiterons à un cas particulier simple où le raisonnement ci-dessus est applicable. Plus précisément :

Théorème 1. Soit g et h deux fonctions continûment dérivables dans un intervalle $[a, b[$ borné ou pas (on suppose cependant a fini) telles que $g(x)\, e^{th(x)}$ soit intégrable sur $[a, b[$ pour t assez grand. Supposons de plus que la fonction h admet un maximum pour $x = a$ tel que $h'(a) = 0$, $h''(a) < 0$, et $g(a) = 0$ et que le maximum de $h(x)$ dans tout sous-intervalle $[a', b[$, avec $a' > a$, est inférieur à $h(a)$; on a alors :

$$\int_a^b g(x)\, e^{th(x)}\, dx \sim g(a)\, e^{th(a)} \sqrt{\frac{-\pi}{2\, t h''(a)}}, \quad t \to \infty.$$

Par exemple, si :

$$I_n(t) = \frac{1}{\pi} \int_0^{\pi} e^{t \cos x} \cos nx\, dx,$$

on obtient :

$$I_n(t) \sim \frac{e^t}{\sqrt{2\pi t}}, \quad t \to \infty.$$

Donnons également une application à la fontion gamma d'Euler :

$$\Gamma(t) = \int_0^{\infty} e^{-u} u^{t-1}\, du,$$

qui extrapole la factorielle, à savoir, pour n entier naturel, $\Gamma(n + 1) = n\,!$. Après changement de variable $u = xt$, on obtient :

$$\Gamma(t + 1) = t^{t+1} \int_0^{\infty} e^{t(-x + \ln x)}\, dx,$$

où l'intégrale est la forme ci-dessus avec $h(x) = \ln x - x$. La fonction h admet un unique maximum pour $x = 1$ dans l'intervalle $(0, +\infty)$ et $h''(1) = -1$; appliquant le théorème 2 à chacun des intervalles $(0, 1)$ et $(1, +\infty)$, ce qui revient à multiplier par 2 la formule du théorème 1, on obtient la célèbre formule de Stirling :

$$\Gamma(t + 1) \sim \sqrt{2\pi t}\, e^{-t}\, t^t, \quad t \to \infty,$$

qui précise le comportement asymptotique de la fonction gamma lorsque t tend vers l'infini. Le développement asymptotique de $\ln \Gamma$ peut être obtenu par la formule d'Euler-Maclaurin.

La méthode de la phase stationnaire

La méthode de la phase stationnaire a été utilisée par lord Kelvin en 1887, à propos de problèmes d'hydrodynamique, pour étudier des intégrales du type :

$$\mathrm{I}(t) = \int_a^b g(x)\, e^{ith(x)}\, dx,$$

où g et h sont des fonctions très régulières dans (a, b) ; dans les applications physiques, $g(x)$ apparaît comme l'amplitude et $th(x)$ comme la phase du phénomène considéré. L'idée essentielle de Stokes et Kelvin est que la partie prédominante de l'intégrale pour t tendant vers l'infini est obtenue au voisinage des extrémités de l'intervalle d'intégration et surtout au voisinage des points c où la phase est « stationnaire », c'est-à-dire tels que $h'(c) = 0$. Intuitivement, lorsque t est grand, le point $g(x)\, e^{ith(x)}$ tourne « très vite » autour de 0 dans le plan complexe au voisinage de tout point c de l'intervalle (a, b) tel que $h'(c) \neq 0$, et cela a pour conséquence de rendre « petite » la partie correspondante de l'intégrale ; si maintenant $h'(c) = 0$, la phase est stationnaire au voisinage de c, c'est-à-dire que cette rotation est très ralentie et la contribution de l'intégrale au voisinage d'un tel point est prédominante sur le reste. Ce principe peut présenter de nombreux aspects ; nous nous limiterons, à titre indicatif, à deux énoncés mettant en évidence le rôle joué par les extrémités de l'intervalle d'intégration, d'une part, et par la présence de points où la phase est stationnaire, d'autre part. En fait, c'est la contribution de ces derniers qui est prépondérante sur la contribution des extrémités.

Théorème 2 (rôle des extrémités de l'intervalle d'intégration). Soit g et h deux fonctions indéfiniment dérivables dans l'intervalle compact $[a, b]$. Si la fonction h n'a pas de point stationnaire dans l'intervalle, on a :

$$\int_a^b g(x)\, e^{ith(x)}\, dx = \frac{g(b)}{it\,h'(b)}\, e^{ith(b)} - \frac{g(a)}{it\,h'(a)}\, e^{ith(a)} + \mathrm{O}\left(\frac{1}{t^2}\right), \quad t \to \infty.$$

Théorème 3 (rôle des points où la phase est stationnaire). Supposant encore que g et h sont indéfiniment dérivables dans l'intervalle compact $[a, b]$, nous supposerons, de plus, que h' s'annule en un seul point c de cet intervalle avec $g(c) \neq 0$ et $h''(c) \neq 0$. Alors on a, pour t tendant vers l'infini :

$$\mathrm{I}(t) = \begin{cases} \sqrt{\dfrac{\pi}{2th''(c)}}\, g(c)\, e^{ith(c)} e^{i\pi/4} + \mathrm{O}\left(\dfrac{1}{t}\right) & \text{si } h''(c) > 0, \\ \sqrt{\dfrac{-\pi}{2th''(c)}}\, g(c)\, e^{ith(c)} e^{-i\pi/4} + \mathrm{O}\left(\dfrac{1}{t}\right) & \text{si } h''(c) < 0, \end{cases}$$

Nous renvoyons aux ouvrages spécialisés pour l'étude détaillée des cas obtenus par superposition des phénomènes ci-dessus et pour le cas des intervalles non compacts.

La méthode du col

Cette méthode a été utilisée par Riemann en 1863 pour étudier le comportement asymptotique de la fonction hypergéométrique et Debye l'a systématiquement développée dans deux mémoires de 1909 et 1910. Il s'agit d'étudier, pour t réel tendant vers l'infini, des intégrales du type :

$$\mathrm{I}(t) = \int_{\mathrm{L}} g(z)\, e^{th(z)}\, dz,$$

où L est un chemin, fini ou pas (pouvant dépendre de t), contenu dans un ouvert D du plan complexe dans lequel g et h sont

des fonction analytiques. Le théorème de Cauchy montre qu'on peut, dans de nombreux cas, modifier le chemin L sans changer la valeur de l'intégrale ; la méthode du col consiste à profiter de cette possibilité en choisissant un chemin L′ sur lequel la fonction $w(z) = |e^{th(z)}| = e^{t\text{Re}h(z)}$ n'atteigne son maximum qu'en un seul point, dans des conditions qui permettent d'appliquer la méthode de Laplace (un peu modifiée pour tenir compte du fait que h prend des valeurs complexes). Intuitivement, on choisira le chemin L′ passant par un point z_0 de telle sorte que, sur L′ au voisinage de ce point, la phase $\text{Im}\, h(z)$ varie peu, alors que l'amplitude $w(z)$ décroît assez vite. Indiquons l'aspect géométrique de cette question, ce qui justifiera la terminologie utilisée. Considérons dans l'espace à trois dimensions la surface d'équation $z = w(x + iy) = e^{t\text{Re}h(x+iy)}$, appelé le *relief* de $e^{th(z)}$. Cette surface ne présente pas de « sommet » relatif, d'après le principe du maximum pour les fonctions analytiques, et, par suite, les seuls points où le plan tangent est horizontal (ce sont les points où la dérivée $h'(z)$ s'annule), sont des cols ; on dira donc qu'un point z_0 du plan complexe tel que $h'(z_0) = 0$ est un *col* ; l'ordre du col est, par définition, le nombre m tel que :

$$h'(z_0) = \dots = h^{(m)}(z_0) = 0,$$
$$h^{(m+1)}(z_0) \neq 0.$$

Revenons à l'interprétation géométrique ; les *lignes de niveaux* du relief (c'est-à-dire les courbes $w = C^{te}$) ont pour projection sur le plan des $z = x + iy$, les courbes sur lesquelles $\text{Re}\,th(z)$ reste constant ; les *lignes de plus grande pente* du relief, qui sont orthogonales aux lignes de niveau, ont pour projection les courbes du plan complexe sur lesquelles la partie imaginaire de $th(z)$ reste constant. Le long d'une telle courbe, $e^{th(z)}$ a une phase constante, alors que son module varie « le plus vite possible », car les lignes de plus grande pente sont les lignes de gradient du relief. Limitons-nous, pour simplifier, au cas d'un col d'ordre 2. Par un tel point passent deux lignes de niveau et deux lignes de plus grande pente, la ligne de faîte, qui suit la crête, et la ligne de thalweg, qui descend dans la vallée. L'idée, ici, est essentiellement de chercher une ligne L′ qui passe par un seul col et qui soit aussi proche que possible (au voisinage de ce col) de la projection de la ligne de thalweg ; en effet, sur cette dernière, la phase de $e^{th(z)}$ reste constante et le module de $e^{th(z)}$ présente un maximum au col. Si on sait trouver une telle ligne, la méthode du col consiste à remplacer l'intégrale étudiée par l'intégrale prise le long d'un petit segment de la tangente à la ligne de thalweg du col, et de remplacer g et h par leurs développements aymptotiques dans cette intégrale.

Donnons un exemple significatif de cette méthode. Dans ce qui suit, nous considérerons seulement des chemins infinis L :

$$s \to r(s)\, e^{i\varphi(s)},$$

définis pour $-\infty < s < +\infty$, et satisfaisant aux conditions (A) :

$$\lim_{s \to +\infty} r(s) = +\infty$$

et :

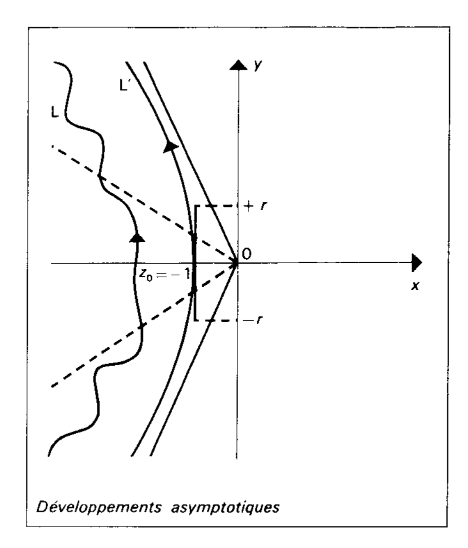
$$\frac{2\pi}{3} \leqslant \varphi(s) \leqslant \frac{2\pi}{3} + \frac{\pi}{6},$$

au voisinage de $+\infty$;

$$\lim_{s \to -\infty} r(s) = +\infty$$

$$-\frac{2\pi}{3} - \frac{\pi}{6} \leqslant \varphi(s) \leqslant -\frac{2\pi}{3},$$

au voisinage de $-\infty$ (cf. figure).



Développements asymptotiques

On voit facilement, en appliquant le théorème de Cauchy et en passant à la limite, que l'intégrale d'Airy, introduite en 1838 par Airy dans ses recherches sur l'optique :

$$\mathrm{Ai}(t^2) = \frac{1}{2i\pi}\int_{\mathrm{L}} \exp(t^2 w - \frac{1}{3}w^3)\,dw$$

ne dépend pas du chemin L, pourvu qu'il satisfasse aux conditions (A). Effectuant le changement de variable : $w = tz$, on a :

$$\mathrm{Ai}(t^2) = \frac{t}{2i\pi}\int_{\mathrm{L}_1} \exp(t^3(z - \frac{1}{3}z^3))\,dz,$$

où le chemin $\mathrm{L}_1 = t^{-1}\mathrm{L}$ satisfait encore (A) ; appliquons la méthode du col à cette intégrale. La fonction $h(z) = z - z^3/3$ a pour dérivée $h'(z) = 1 - z^2$, d'où deux cols $z_0 = +1$ et $z_0 = -1$. Ici, la ligne de faîte passant par z_0 est l'axe réel, et la ligne de thalweg est la branche L′ de l'hyperbole $1 - x^2 + y^2/3 = 0$ telle que $x < 0$ (cf. figure). Ce chemin satisfait aux conditions (A) et on peut ici prendre exactement la ligne de thalweg comme nouveau chemin d'intégration sans changer la valeur de l'intégrale. Remplaçant alors L′ au voisinage du col $z_0 = -1$ par le segment tangent $u \rightarrow -1 + iu, -r \leqslant u \leqslant +r$, et $h(-1 + iu)$ par les deux premiers termes de son développement de Taylor $-2/3 - u^2$, on obtient l'intégrale :

$$\frac{te^{-2t^3/3}}{2\pi}\int_{-r}^{+r} e^{-t^3u^2}du.$$

La formule de Laplace, appliquée à chacun des deux intervalles $[-r, 0]$ et $[0, r]$ donne :

$$\int_{-r}^{+r} e^{-t^3u^2}\,du \sim 2\sqrt{\frac{\pi}{4t^3}}, \quad t \rightarrow \infty;$$

d'où finalement :

$$\mathrm{Ai}(t^2) \sim \frac{e^{-2t^3/3}}{2\sqrt{\pi t}}, \quad t \rightarrow \infty.$$

6. Cas des solutions d'équations différentielles

On se bornera aux systèmes d'équations différentielles linéaires, en distinguant deux cas, le champ réel et le champ complexe.

Systèmes dans le champ réel

Plaçons-nous d'abord dans le cas d'un système linéaire à coefficients constants :

(1) $$x' = Ax,$$

où A est une matrice carrée d'ordre n à coefficients complexes et $x : t \mapsto x(t)$ une fonction de classe $\mathcal{C}^1$ sur $[0, +\infty[$ à valeurs dans $\mathbf{C}^n$. Pour toute condition initiale $a \in \mathbf{C}^n$, l'unique solution du problème de Cauchy $x(0) = a$ est donnée par :

$$x(t) = (\exp tA)\,a$$

(cf. équations DIFFÉRENTIELLES). Lorsque A est diagonalisable, de valeurs propres

λ_1 ...., λ_r, le comportement asymptotique de $x(t)$ est gouverné par la valeur propre de plus grande partie réelle. En particulier, les solutions tendent vers 0 à l'infini si et seulement si, pour tout j, on a $\operatorname{Re}\lambda_j \leqslant 0$. Lorsque A n'est pas diagonalisable et que λ_j est d'indice n_j, il existe des solutions se comportant comme $t^k e^{jt}$, où $0 \leqslant k \leqslant n_j - 1$. Les solutions tendent encore vers 0 à l'infini si et seulement si $\operatorname{Re}\lambda_j < 0$.

Examinons maintenant l'effet d'une perturbation $t \mapsto R(t)$ de A sur le comportement asymptotique d'une solution du système linéaire (1). On peut conjecturer que, si la perturbation est assez petite à l'infini, ce comportement n'est pas notablement modifié. Plus précisément, supposons A diagonalisable et soit λ une valeur propre de A. Si l'intégrale :

$$\int_1^\infty \|R(t)\| dt$$

est convergente, alors, pour tout vecteur propre b de A associé à la valeur propre λ, il existe une solution x et une seule de l'équation perturbée :

$$x' = (A + R)x$$

telle que $x(t) \sim e^{\lambda t}$ au voisinage de $+\infty$.

Par exemple, soit l'équation de Bessel réduite (cf. fonctions de BESSEL, pour ce qui suit) :

$$u'' + \left(1 - \frac{\nu^2}{t^2}\right)u = 0,$$

qui équivaut au système linéaire :

$$\begin{cases} x_1 = x_2 \\ x_2 = -\left(1 - \frac{\nu^2}{t^2}\right)x_1 ; \end{cases}$$

ici :

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad R(t) = \begin{pmatrix} 0 & 0 \\ \gamma^2/t^2 & 0 \end{pmatrix}.$$

Les fonctions $t \mapsto e^{it}$, et $t \mapsto e^{-it}$ constituent une base de l'espace vectoriel des solutions de l'équation non perturbée $u'' + u = 0$. D'après le résultat précédent, il existe donc un couple (u_1, u_2) de solutions et un seul de l'équation perturbée tel que :

$$u_1(t) \sim e^{it}, \quad u_2(t) \sim e^{-it} ;$$

cette méthode donne donc le comportement asymptotique des fonctions de Bessel.

Si, maintenant, A n'est pas diagonalisable, il convient d'imposer à $\mathbf{R}$ des conditions plus strictes. Lorsque λ_j est d'indice n_j, on suppose que :

$$\int_1^\infty t^{n_j - 1} \|\mathbf{R}(t)\| dt < +\infty,$$

ce qui entraîne l'existence d'une solution x telle que :

$$x(t) \sim t^{n_j - 1} e^{\lambda_j t} b.$$

Pour le comportement asymptotique des systèmes linéaires à cœfficients périodiques (par exemple, le théorème de l'équation séculaire des planètes), on se reportera à l'article équations DIFFÉRENTIELLES, *La théorie de Floquet*.

Le champ complexe

Pour obtenir des développements asymptotiques des solutions d'un système différentiel dans le champ complexe, l'idée principale consiste à obtenir des représentations intégrales des solutions. On utilise ensuite des développements tayloriens ou on applique à ces intégrales les méthodes esquissées au chapitre 5.

Considérons, par exemple, une équation différentielle linéaire d'ordre n :

$$a_0(z)u^{(n)} + a_1(z)u^{(n-1)} + \dots + a_n(z)u = 0,$$

où $a_0, a_1 \dots, a_n$ sont des polynômes.

Pour obtenir des représentations intégrales des solutions, les méthodes sont très variées. Citons, par exemple, la méthode de Laplace, qui s'applique lorsque les polynômes a_j sont de degré $\leqslant 1$: on cherche les solutions sous la forme :

$$u(z) = \int_L v(\zeta)e^{\zeta z}d\zeta,$$

où L est un contour du plan complexe convenablement choisi.
Ainsi, dans le cas de l'équation différentielle :

$$u'' - zu = 0$$

on trouve $v(z) = \exp(-\zeta^3/3)$ et :

$$u(\zeta) = \int_L \exp(\zeta z - \frac{1}{3}\zeta^3)\,d\zeta,$$

où L est le contour décrit dans l'intégrale d'Airy (cf. *La méthode du col*, in chap. 5). Le comportement asymptotique de la solution s'en déduit.

On utilise aussi la méthode d'Euler, où :

$$u(z) = \int_L v(\zeta)(\zeta - z)^\alpha d\zeta,$$

pour un choix convenable de α et L. Une variante est la méthode de Mellin, où :

$$u(z) = \int_L v(\zeta)(1 - \zeta z)^\alpha d\zeta.$$

Nous nous bornerons à deux exemples significatifs, importants pour les applications.

L'équation hypergéométrique

Considérons l'équation de Riemann (cf. *Les équations différentielles de la physique mathématique*, in chap. 2 de équations DIFFÉRENTIELLES), admettant trois points singuliers deux à deux distincts. Par une transformation homographique, cette équation se ramène à l'équation hypergéométrique :

$$z(1-z)\frac{d^2u}{dz^2} + \{c - (a+b+1)z\}\frac{du}{dz} - abu = 0,$$

où a, b, c sont des nombres complexes. Lorsque c n'est pas un entier négatif, on obtient une solution holomorphe dans le disque $|z| < 1$ par la série entière, dite hypergéométrique :

$$F(a,b,c,z) = 1 + \frac{a.b}{1.c}z + \frac{a(a+1)b(b+1)}{1.2.c(c+1)}z^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{1.2.3.c(c+1)(c+2)}z^3 + \dots$$

qui s'écrit aussi :

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)}F(a,b,c,z) = \sum_{n=0}^{\infty}\frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(n+1)\Gamma(c+n)}z^n,$$

où Γ désigne la fonction gamma d'Euler (cf. fonction GAMMA).

On peut prolonger analytiquement $F(a, b, c, z)$ au plan fendu $\mathbf{C} - \mathbf{R}^+$ par la représentation intégrale de Mellin-Barnes :

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)}F(a,b,c,z) = \frac{1}{2i\pi}\int_D \frac{\Gamma(a+\zeta)\Gamma(b+\zeta)\Gamma(-\zeta)}{\Gamma(c+\zeta)}(-z)^\zeta d\zeta,$$

où D désigne l'axe imaginaire et $(-z)^\zeta = \exp(\zeta \ln(-z))$ en prenant la détermination principale du logarithme, c'est-à-dire $|\arg(-z))| < \pi$.

Cette représentation intégrale permet alors d'obtenir la relation suivante :

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(c)}F(a,b,c,z) = \frac{\Gamma(a)\Gamma(a-b)}{\Gamma(a-c)}(-z)^{-a}F(a, 1-c+a; 1-b+a; z^{-1}) + \frac{\Gamma(b)\Gamma(b-a)}{\Gamma(b-c)}(-z)^{-b}F(b, 1-c+b; 1-a+b; z^{-1}),$$

qui fournit aussitôt le développement asymptotique de $F(a, b, c, z)$ au voisinage de l'infini sur $\mathbf{C} - \mathbf{R}^+$.

L'équation hypergéométrique confluente

L'équation hypergéométrique confluente est le cas où deux des trois singularités de l'équation de Riemann confluent en un même point. Ici encore, on se ramène au cas où ce point est à l'infini et où l'autre singularité est au point 0. On obtient alors l'équation de Whittaker, qui s'écrit :

$$\frac{d^2u}{dz^2} + \left(-\frac{1}{4} + \frac{k}{z} + \frac{1-4m^2}{4z^2}\right)u = 0,$$

où k et m sont des nombres complexes. La méthode de Laplace ou celle de Mellin fournissent une solution de la forme :

$$W_{k,m}(z) = \frac{e^{-\frac{1}{2}z}z^k}{\Gamma(\frac{1}{2} - k + m)} \times \int_0^\infty t^{-k-\frac{1}{2}+m}\left(1 + \frac{t}{z}\right)^{k-\frac{1}{2}+m} e^{-t}dt.$$

sous condition que $\mathrm{Re}(k - \frac{1}{2} - m) \leqslant 0$, qui est holomorphe dans le plan fendu $\mathbf{C} - \mathbf{R}$. Cette fonction est appelée fonction de Whittaker. En développant :

$$\left(1 + \frac{t}{z}\right)^{k-\frac{1}{2}+m}$$

en série de Taylor, on obtient le développement asymptotique de $W_{k,m}$ au voisinage de l'infini sur $\mathbf{C} - \mathbf{R}^-$.

De nombreuses fonctions classiques apparaissent comme des cas particuliers des fonctions de Whittaker. Ainsi, la fonction d'erreur :

$$\mathrm{Er}f(x) = 1 - \frac{1}{\sqrt{\pi x}}\exp\left(-\frac{x^2}{2}\right)W_{-\frac{1}{4},\frac{1}{4}}(x^2).$$

Il en est de même pour le logarithme intégral. Enfin, l'équation de Bessel réduite :

$$u'' + \left(-\frac{1}{4} + \frac{1-4\nu^2}{4z^2}\right)u = 0,$$

est satisfaite par $W_{0,n}(z)$ et $W_{0,n}(-z)$. La décomposition de la fonction de Bessel J_ν dans cette base (cf. fonctions de BESSEL) s'écrit :

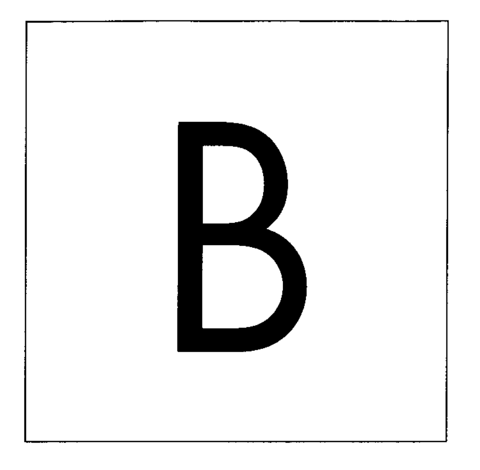
$$J_\nu(z) = \frac{1}{\sqrt{2\pi z}}\left[W_{0,\nu}(2iz)\exp\left(\frac{1}{2}(\nu + \frac{1}{2})\pi i\right) + W_{0,\nu}(-2iz)\exp\left(-\frac{1}{2}(\nu + \frac{1}{2})\pi i\right)\right],$$

ce qui fournit un développement de J_ν à toute précision.

JEAN-LOUIS OVAERT et JEAN-LUC VERLEY

Bibliographie

N. G. DE BRUIJN, *Asymptotic Methods in Analysis*, Dover, 1982 / E. COPSON, *Asymptotic Expansions*, Cambridge Univ. Press, Cambridge, 1965 / J. DIEUDONNÉ, *Calcul infinitésimal*, Hermann, 2ᵉ éd. 1980 / A. ERDÉLYI, *Asymptotic Expansions*, Dover Publications, New York, 1961 / G. H. HARDY, *Orders of Infinity*, Cambridge Univ. Press, 1910 / E. INCE, *Ordinary Differential Equations*, Dover Publications, New York, 1956 / B. R. VAINBERG, *Asymptotic Methods in Equations of Mathematical Physics*, Gordon & Breach, New York, 1989 / E. T. WHITTAKER & G. N. WATSON, *A Course of Modern Analysis*, 4ᵉ éd., 1927, Cambridge Univ. Press, réimpr. 1969.



BARYCENTRE

Soit A un espace affine attaché à un espace vectoriel E (sur un corps commutatif K). On appelle « point M de A affecté de la masse λ » l'élément (M, λ) de l'ensemble A × K.

Par définition, le barycentre de n points $M_1, M_2, ..., M_n$ de A affectés des masses $\lambda_1, \lambda_2, ..., \lambda_n$ de somme non nulle est le point G tel que :

$$\sum_{i=1}^{n} \lambda_i \mathbf{GM}_i = 0.$$

De cette définition découlent aisément plusieurs propriétés :

1. Pour tout point O de A, on a la relation (équivalente à la condition de définition) :

$$\left(\sum_{i=1}^{n} \lambda_i\right) \mathbf{OG} = \sum_{i=1}^{n} \lambda_i \mathbf{OM}_i.$$

2. Le barycentre de la famille (M_i, λ_i) est le même que le barycentre de la famille des $(M_i, \alpha\lambda_i)$, où α est un élément non nul de K.

3. Propriété d'associativité : soit G le barycentre de n points $M_1, M_2, ..., M_n$ de A affectés des masses $\lambda_1, \lambda_2, ..., \lambda_n$ et soit G′ le barycentre des points $M_1, M_2, ..., M_k$ affectés des masses $\lambda_1, \lambda_2, ..., \lambda_k$. Alors G est aussi le barycentre des points $G', M_{k+1}, ..., M_n$ affectés des masses :

$$\sum_{i=1}^{k} \lambda_i, \lambda_{k+1}, \lambda_n.$$

Lorsque le corps K est de caractéristique 0 et que les scalaires λ_i sont égaux, le barycentre G s'appelle centre de gravité, ou équibarycentre de la famille des (M_i, λ_i).

JACQUES MEYER

BESSEL FONCTIONS DE

Les fonctions de Bessel jouent un rôle important en mathématiques appliquées et en physique mathématique. Elles interviennent aussi bien dans des problèmes de conduction de la chaleur que dans des problèmes de diffraction, acoustique, ou électromagnétisme. Elles apparaissent souvent dans l'étude d'équations différentielles ou d'équations aux dérivées partielles avec des conditions aux limites relatives à des frontières sphériques ou cylindriques. Ces fonctions possèdent certaines analogies avec les fonctions trigonométriques : comportement à l'infini, zéros, développement en série.



Définitions

Considérons l'équation différentielle :

$$(1) \qquad y'' + \frac{1}{x}y' + \left(1 - \frac{\nu^2}{x^2}\right)y = 0,$$

dans laquelle ν est un paramètre complexe quelconque. Les fonctions de Bessel, ainsi que d'autres fonctions voisines, les fonctions de Neumann et de Hankel, sont des solutions particulières de cette équation différentielle. Supposant que x est une variable réelle positive, cherchons des solutions de la forme :

$$y = x^{\alpha} \sum_{k=0}^{\infty} c_k x^k.$$

Par identification terme à terme, on obtient :

$$(\alpha^2 - \nu^2)\, c_0 = 0,$$
$$[(1 + \alpha)^2 - \nu^2]\, c_1 = 0,$$
$$[(k + \alpha)^2 - \nu^2]\, c_k + c_{k-2} = 0, \quad k \geqslant 2 ;$$

donc, on doit avoir $\alpha = \pm\, \nu$.

Si on choisit $\alpha = \nu$, on trouve :

$$c_{2k} = (-1)^k \frac{c_0}{2^{2k}\, k\,!\, (\nu + 1)\ldots(\nu + k)} ;$$

on est alors amené à choisir :

$$c_0 = \frac{1}{2^{\nu}\, \Gamma(\nu + 1)},$$

d'où :

$$c_{2k} = (-1)^k \frac{1}{2^{2k}\, k\,!\, \Gamma(k + \nu + 1)} ;$$

Γ désigne ici la fonction eulérienne ; c'est une fonction continue qui a pour propriété fondamentale $\Gamma(n + 1) = n\,!$ pour n entier (cf. fonction GAMMA).

Par définition, la fonction de Bessel J_{ν}, d'indice ν, est donnée par ces coefficients, soit :

$$J_{\nu}(x) = \left(\frac{x}{2}\right)^{\nu} \sum_{k=0}^{\infty} \frac{(-1)^k}{k\,!\, \Gamma(k + \nu + 1)} \left(\frac{x}{2}\right)^{2k}.$$

La série qui intervient dans la définition de J_{ν} converge pour tout x réel et on vérifie que J_{ν} satisfait à l'équation différentielle donnée. Si ν est entier, J_{ν} peut être définie pour tout x complexe. Si ν n'est pas entier, on peut étendre le domaine de définition de J_{ν} à l'ensemble des x complexes non nuls, mais l'origine sera un point critique à cause du facteur $(x/2)^{\nu}$.

Si ν n'est pas entier, la solution $\alpha = -\nu$ fournit la fonction :

$$J_{-\nu}(x) = \left(\frac{x}{2}\right)^{-\nu} \sum_{k=0}^{\infty} \frac{(-1)^k}{k\,!\, \Gamma(k - \nu + 1)} \left(\frac{x}{2}\right)^{2k}.$$

$J_{-\nu}$ est encore une solution de l'équation différentielle (1), et on vérifie que $J_{-\nu}$ et J_{ν} sont linéairement indépendantes. La solution générale de (1) est donc :

$$\alpha J_{\nu}(x) + \beta J_{-\nu}(x),$$

α et β étant deux constantes arbitraires. Si ν est un entier n, on constate que :

$$J_{-n}(x) = (-1)^n J_n(x).$$

Tous ces faits peuvent être justifiés en remarquant que l'équation différentielle (1) est une équation du type de Fuchs (cf. équations DIFFÉRENTIELLES), dont l'équation déterminante est $x^2 + \nu^2 = 0$. La différence des racines de l'équation déterminante est 2ν. Si 2ν n'est pas entier, on a une base de la forme $x^{\nu}\, A(x)$ et $x^{-\nu}\, B(x)$, $A(x)$ et $B(x)$ étant deux fonctions holomorphes au voisinage de $x = 0$. Si 2ν est entier, on n'obtient ainsi qu'une seule solution $x^{\nu}\, A(x)$. Pour obtenir une deuxième solution dans ce cas, on est amené à introduire les fonctions de Neumann. On appelle fonction de Neumann d'indice ν, N_{ν}, la fonction :

$$N_{\nu}(x) = \frac{\cos \nu\pi\, J_{\nu}(x) - J_{-\nu}(x)}{\sin \nu\pi}$$

pour ν non entier et :

$$N_n(x) = \lim_{\varepsilon \to 0} N_{n+\varepsilon}(x),$$

pour n entier. On vérifie alors que, pour tout ν, la fonction de Neumann est une solution de (1) et que N_n et J_n sont linéairement indépendantes. Donc, si n est entier, la solution générale de (1) est :

$$\alpha J_n + \beta N_n.$$

Enfin, on introduit parfois les fonctions de Hankel :

$$H^1_\nu = J_\nu + i N_\nu,$$
$$H^2_\nu = J_\nu - i N_\nu ;$$

que ν soit entier ou non, H^1_ν et H^2_ν forment une base de solutions de (1).

Propriétés principales

Formules asymptotiques. Le comportement à l'infini des fonctions introduites est donné par les formules :

$$J_\nu(x) = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + \varphi_1(x),$$

$$N_\nu(x) = \sqrt{\frac{2}{\pi x}} \sin\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + \varphi_2(x),$$

$$H^1_\nu(x) = \sqrt{\frac{2}{\pi x}} \exp\left[i\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right)\right] + \varphi_3(x),$$

$$H^2_\nu(x) = \sqrt{\frac{2}{\pi x}} \exp\left[-i\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right)\right] + \varphi_4(x),$$

où les fonctions $\varphi_i(x)$ tendent vers zéro lorsque x tend vers l'infini avec $|\arg x| \leqslant \pi - \varepsilon$, $\varepsilon > 0$; d'une manière plus précise, les quantités $x^{3/2}\varphi_i(x)$ restent bornées.

Formules de récurrence. À partir des développements en séries, on obtient les deux formules de récurrence :

$$J_{\nu-1}(x) + J_{\nu+1}(x) = \frac{2\nu}{x} J_\nu(x),$$
$$J_{\nu-1}(x) - J_{\nu+1}(x) = 2J'_\nu(x).$$

Voici une application de ces formules.

On vérifie directement que :

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}} \sin x,$$

$$J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}} \cos x,$$

par suite, si ν est de la forme $n + 1/2$, avec n entier, les formules de récurrence montrent que J_ν est une combinaison linéaire de fonctions élémentaires.

Représentation intégrale. Il est souvent commode de représenter J_ν sous la forme d'une intégrale. On a la formule :

$$J_\nu(x) = \frac{(x/2)^\nu}{\Gamma(\nu + 1/2)\sqrt{\pi}} \int_{-1}^{+1} (1 - t^2)^{\nu - 1/2} e^{-itx} dt,$$

qui est valable pour $\operatorname{Re} \nu + 1/2 > 0$.

Fonction génératrice. Les fonctions de Bessel sont liées à un développement en série de Laurent. En effet, introduisons la fonction :

$$F(u) = \exp\left(\frac{u}{2} - \frac{1}{2u}\right) x,$$

pour u complexe et x paramètre complexe non nul ; son développement en série de Laurent à l'origine s'écrit :

$$F(u) = \sum_{-\infty}^{+\infty} u^n J_n(x)$$

(formule de Schlömilch). La fonction $F(u)$ est appelée fonction génératrice. Changeant u en $1/u$ et faisant le produit membre à membre des égalités obtenues, on obtient :

$$1 = [J_0(x)]^2 + 2 \sum_{1}^{\infty} [J_n(x)]^2 ;$$

on en déduit, pour tout x réel, les majorations :

$$|J_0(x)| < 1, \quad |J_n(x)| < \frac{1}{\sqrt{2}}, \quad n \geqslant 1.$$

Voici une autre présentation de la formule de Schlömilch ; posant $u = e^{i\theta}$, on obtient :

$$\exp(ix \sin \theta) = \sum_{-\infty}^{+\infty} e^{in\theta} J_n(x) ;$$

les $J_n(x)$ sont donc ainsi les coefficients de Fourier de la fonction :

$$\theta \mapsto \exp(ix \sin \theta).$$

Théorème de Neumann. À partir des fonctions de Bessel on peut obtenir des représentations analogues au développement en séries entières. Plus précisément, toute fonction holomorphe dans le disque $|z| < R$ peut être développée en série sous la forme :

$$f(z) = \sum_{n=0}^{\infty} a_n J_n(z),$$

où la série converge absolument et uniformément dans tout disque compact $|z| \leqslant R_1 < R$.

De plus, il faut signaler que l'on peut faire une théorie analogue à la transformation de Fourier en employant les fonctions de Bessel : soit $f(x)$ une fonction définie pour $x \geqslant 0$, à valeurs réelles ou complexes, continue par morceaux, telle que :

$$\int_0^{\infty} x |f(x)| \, dx < \infty ;$$

pour tout entier n, positif, posons :

$$g_n(s) = \int_0^{\infty} t f(t) J_n(st) \, dt ;$$

alors, on a la formule suivante d'inversion :

$$f(x) = \int_0^{\infty} u J_n(xu) g_n(u) \, du.$$

Des fonctions voisines des fonctions de Bessel sont parfois utilisées. Ce sont, par exemple :

$$I_\nu(x) = \exp\left(-\frac{i\nu\pi}{2}\right) J_\nu(ix),$$

$$\rho_\nu(x) = \sqrt{\frac{\pi}{2x}} \times J_{\nu+\frac{1}{2}}(x),$$

appelées respectivement fonctions de Bessel modifiées, et fonctions de Bessel sphériques.

PIERRE SAPHAR

Bibliographie

R. Courant & D. Hilbert, *Methods of Mathematical Physics*, 2 vol., Wiley, New York, 1989 / E. Grosswald, « Bessel Polynomials », in *Lecture Notes in Mathematics*, Springer-Verlag, New York, 1979 / F. Oberhettinger, *Tables of Bessel Transforms*, *ibid.*, 1972.

BOOLE ALGÈBRE & ANNEAU DE

La notion d'algèbre de Boole, introduite par G. Boole (1847) et par A. De Morgan afin d'algébriser les opérations propositionnelles de la logique, joue un rôle très utile dans plusieurs branches des mathématiques (algèbre, théorie des ensembles ordonnés, calcul des probabilités) et en logique mathématique (logique algébrique, modèles booléens).

On appelle *algèbre de Boole* (A, $\vee$, $\wedge$, $\neg$, 0, 1) la donnée d'un ensemble A (non vide) muni de deux lois de composition interne $\vee$ et $\wedge$, associatives et commutatives, d'une application unaire $\neg$ et de deux éléments privilégiés 0 et 1, ces données vérifiant les axiomes suivants :

$$\forall x \in A \ \forall y \in A \ (x \wedge y) \vee y = y;$$
$$\forall x \in A \ \forall y \in A \ (x \vee y) \wedge y = y;$$
$$\forall x \in A \ \forall y \in A \ \forall z \in A x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z);$$
$$\forall x \in A \ \forall y \in A \ \forall z \in A x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z);$$
$$\forall x \in A \ x \wedge \neg x = 0;$$
$$\forall x \in A \ x \vee \neg x = 1;$$
$$\forall x \in A \ 0 \vee x = x;$$
$$\forall x \in A \ 1 \wedge x = x;$$

On appelle *anneau de Boole* la donnée (A ; +, ., 0, 1) d'un anneau commutatif unitaire vérifiant :

$$\forall x \in A \quad x^2 = x.$$

Les structures d'algèbre de Boole et d'anneau de Boole sont équivalentes au sens suivant :

– On peut associer à toute algèbre de Boole (A, $\vee$, $\wedge$, $\neg$, 0, 1) l'anneau de Boole (A ; +, ., 0, 1) défini par :

$$\forall x \in A \ \forall y \in A \ x + y = (x \wedge \neg y) \vee (y \wedge \neg x),$$
$$\forall x \in A \ \forall y \in A \ x . y = x \wedge y;$$

– On peut associer à tout anneau de Boole (A ; +, ., 0, 1) l'algèbre de Boole (A, $\vee$, $\wedge$, $\neg$, 0, 1) définie par :

$$\forall x \in A \ \forall y \in A \ x \vee y = x + y + x . y,$$
$$\forall x \in A \ \forall y \in A \ x \wedge y = x . y$$
$$\forall x \in A \ \neg x = 1 + x.$$

Les deux correspondances précédentes sont inverses l'une de l'autre, comme on le vérifie facilement, et permettent de rattacher la théorie des algèbres de Boole à la théorie des anneaux. On peut également rattacher la théorie des algèbres de Boole à celle des ensembles ordonnés en observant que l'on peut définir un ordre canonique sur toute algèbre de Boole en posant :

$$x \leqslant y \Leftrightarrow x \vee y = y.$$

Exemples d'algèbre de Boole

1. Pour tout ensemble X, l'ensemble $\mathcal{P}$(X) des parties de X devient une algèbre de Boole si on pose :

$$\forall Y \subset X \ \forall Z \subset X \ Y \vee Z = Y \cup Z,$$
$$\forall Y \subset X \ \forall Z \subset X \ Y \wedge Z = Y \cap Z,$$
$$\forall Y \subset X \ \neg Y = X - Y,$$
$$0 = \emptyset, \ 1 = X.$$

Il résulte d'un théorème fondamental, dû à M. Stone, que toute algèbre de Boole est isomorphe à une sous-algèbre de Boole d'une algèbre du type précédent.

2. Soit $\mathcal{F}$ l'ensemble des formules propositionnelles construites à l'aide des connecteurs $\vee$, $\wedge$, $\neg$ à partir d'un ensemble P non vide de variables propositionnelles. Posons A ~ B si et seulement si la formule A $\leftrightarrow$ B est une tautologie. La relation ~ est une relation d'équivalence sur $\mathcal{F}$ compatible avec les connecteurs $\vee$, $\wedge$, $\neg$, ce qui permet de définir sur $\mathcal{F}/\sim$ une structure d'algèbre de Boole. Cette algèbre de Boole est appelée *algèbre de Lindenbaum* du calcul propositionnel P et semble avoir été considérée implicitement par Boole.

Les algèbres de Boole ont servi de prototype au développement de plusieurs techniques. Citons la *logique algébrique* : on adapte l'exemple 2 au contexte plus général de la logique du premier ordre, où l'on doit tenir compte de l'action des quantificateurs ; on obtient ainsi les structures d'algèbres polyadique et cylindrique.

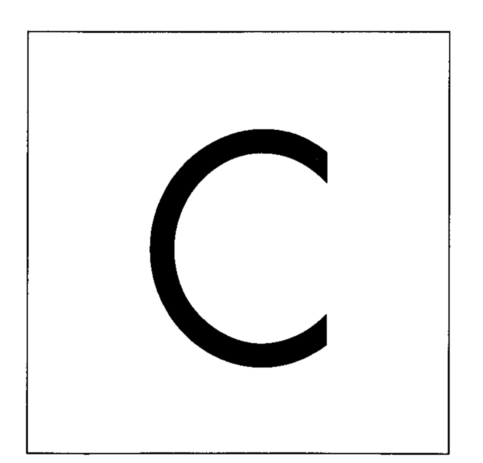
La notion de spectre d'anneau

Le théorème de Stone établit une dualité entre la catégorie des algèbres de Boole et celle des espaces topologiques compacts

totalement discontinus. Plus généralement, la notion de spectre d'anneau fournit un foncteur très utile de la catégorie des anneaux commutatifs dans la catégorie des anneaux topologiques.

Les algèbres de Boole sont d'un emploi constant et traditionnel en théorie de la mesure et en calcul des probabilités. On a introduit récemment avec succès la notion de modèle booléen, qui a permis de donner des démonstrations relativement simples de l'indépendance de l'hypothèse du continu et de faire faire des progrès à la théorie proprement dite des algèbres de Boole (construction d'algèbres de Boole sophistiquées n'ayant aucun automorphisme non trivial, etc.).

GABRIEL SABBAGH



CALCUL DES VARIATIONS → VARIATIONS CALCUL DES

CALCUL INFINITÉSIMAL

A. Calcul à une variable

Créée au XVIIe siècle par Newton, Leibniz et leurs prédécesseurs immédiats, transformée au XVIIIe, par Euler, en un prodigieux instrument de calcul, débarrassée, sous la Restauration, de sa métaphysique par le baron Cauchy, l'analyse infinitésimale a, depuis longtemps, atteint un degré de perfection tel qu'il est devenu possible d'en exposer l'essentiel en moins d'une dizaine de pages. C'est ce que nous allons essayer de faire, en renvoyant le lecteur à l'article qui précède pour des considérations historiques moins schématiques, et en nous plaçant ici au point de vue le plus « unidimensionnel » possible. Le lecteur qui désirerait un exposé plus philosophique et plus historique ne saurait mieux faire que de consulter l'ouvrage classique d'Otto Toeplitz. On n'a voulu, ici, exposer que les résultats les plus importants et les plus simples de la théorie classique à une variable, en s'efforçant de tout démontrer, et en ne demandant du lecteur que les connaissances les plus élémentaires sur les inégalités entre nombres décimaux, plus tout de même, cela va sans dire, une certaine habitude des raisonnements mathématiques.

1. Notion de borne supérieure

Nous désignerons par **R** l'ensemble des *nombres réels* ; il nous suffira de savoir qu'un nombre réel est un développement décimal illimité précédé d'un signe (qu'on omet s'il s'agit du signe +), par exemple le nombre − 3,141 59. ... ou bien le nom-

bre $1 = 1{,}000\,0\ldots\ldots = 0{,}999\,99.\ldots$, et que l'on peut effectuer sur ces nombres des opérations algébriques que tout le monde connaît. On peut aussi comparer deux nombres réels x et y, autrement dit donner un sens à la relation $x < y$ (qui exclut, notons-le, l'égalité $x = y$). On peut, à partir de là, définir des *intervalles* de plusieurs natures ; par exemple, si a et b sont deux nombres réels donnés, on définit quatre intervalles dont a et b sont les extrémités, et qui ne diffèrent entre eux que dans la mesure où ils contiennent, ou non, leurs extrémités : l'intervalle $[a, b]$ est l'ensemble des nombres x tels que $a \leqslant x \leqslant b$, l'intervalle $[a, b[$ est formé des x tels que $a \leqslant x < b$, etc. Les intervalles de la forme $[a, b]$ sont dits *compacts*, et les intervalles de la forme $]a, b[$ sont dits *ouverts*.

Considérons maintenant un ensemble E de nombres réels. On dit qu'il est *borné supérieurement* s'il existe un nombre réel M tel que l'on ait $x \leqslant$ M pour tout $x \in$ E (rappelons que cette notation signifie que le nombre x appartient à E, ou est un élément de E), et borné inférieurement s'il existe un nombre m tel que l'on ait $m \leqslant x$ pour tout $x \in$ E. Si E est borné supérieurement et inférieurement (c'est-à-dire s'il existe un intervalle qui contient E), on dit que E est *borné* tout court. Par exemple, l'ensemble **N** des entiers naturels (ses éléments sont 0, 1, 2, ...) est borné inférieurement, mais non supérieurement, tandis que l'ensemble des nombres rationnels x tels que $x^3 < 2$ est borné supérieurement (en effet $x^3 < 2$ implique $x^3 \leqslant 8 = 2^3$, d'où $x \leqslant 2$, comme on le voit facilement).

Soit E un ensemble borné supérieurement, et soit M un nombre tel que $x \leqslant$ M pour tout $x \in$ E. S'il existe un nombre M' < M tel que l'on ait aussi $x \leqslant$ M' pour tout $x \in$ E, on obtient des informations plus précises sur les éléments de E en écrivant qu'ils sont tous inférieurs à M', qu'en écrivant qu'ils sont tous inférieurs à M (exemple concret : savoir que tout homme vit au plus 500 ans est mieux que rien, mais il vaut mieux savoir que tout le monde meurt avant 200 ans). Pour obtenir, de ce point de vue, les informations les plus précises possibles sur E, on est donc amené à choisir le nombre M aussi petit que possible, d'où la définition suivante : on dit que M est la *borne supérieure* de E si $x \leqslant$ M pour tout $x \in$ E, et si, de plus, il n'existe aucun nombre M' < M possédant la première propriété, ou encore si, pour tout M' < M, il y a au moins un $x \in$ E tel que M' $< x \leqslant$ M. Par exemple, la borne supérieure de l'intervalle $[0, 1[$ est le nombre 1 : on a $x \leqslant 1$ dès que $0 \leqslant x < 1$, et, pour tout M' < 1, il y a des nombres x tels que l'on ait à la fois $0 \leqslant x < 1$ et M' $< x$. On désigne par sup(E) la borne supérieure d'un ensemble E de nombres réels, et par inf(E) sa borne inférieure, définie de façon analogue en renversant les inégalités.

Théorème 1. Tout ensemble non vide borné supérieurement de nombres réels possède une borne supérieure.

La démonstration très simple de ce théorème d'existence (il ne suffit pas de parler d'un objet possédant des propriétés données pour que l'objet en question existe) procède comme suit. Supposons, pour fixer les idées, que l'ensemble E considéré soit contenu dans l'intervalle $[0, 1[$; on notera $0{,}x_1x_2x_3\ldots$ le développement décimal illimité de tout $x \in$ E, de sorte que les chiffres x_k sont des entiers compris entre 0 et 9. Nous allons construire les décimales successives $a_1, a_2, \ldots$ de la borne supérieure cherchée a. On prend pour a_1 la plus grande valeur prise par la décimale x_1 de x lorsque x décrit E (autrement dit : on a $x_1 \leqslant a_1$ pour tout

$x \in E$, et il existe un $x \in E$ tel que $x_1 = a_1$) ; soit alors E_1 l'ensemble des $x \in E$ tels que $x_1 = a_1$; on prend pour a_2 la plus grande valeur prise par la décimale x_2 lorsque x décrit E_1 ; soit alors E_2 l'ensemble des $x \in E_1$ tels que $x_2 = a_2$ (c'est-à-dire des $x \in E$ dont les deux premières décimales sont a_1 et a_2) ; on prend pour a_3 la plus grande valeur prise par x_3 lorsque x décrit E_2, et ainsi de suite indéfiniment. Considérons alors le nombre $a = 0, a_1a_2a_3...$ ainsi construit. On a la relation $x \leqslant a$ pour tout $x \in E$, car si l'on a $x_1 = a_1, ..., x_p = a_p$ (c'est-à-dire si x appartient à l'ensemble E_p construit plus haut), on a $x_{p+1} \leqslant a_{p+1}$ par définition même de a_{p+1} ; on aboutit bien ainsi à la règle classique pour comparer deux développements décimaux. De plus, pour tout entier p, il y a effectivement des $x \in E$ tels que l'on ait $x_1 = a_1, ..., x_p = a_p$, et donc $a - 10^{-p} \leqslant x \leqslant a$; comme il existe, pour tout $M' < a$, un entier p tel que $M' < a - 10^{-p}$, il existe donc a fortiori un $x \in E$ tel que $M' < x$.

Le théorème 1 et des énoncés analogues expriment toute la « métaphysique » du calcul infinitésimal, à savoir l'existence d'un nombre réel possédant un développement décimal arbitrairement donné. Toutes ces constructions s'écrouleraient si l'on ne connaissait que les nombres rationnels, car la borne supérieure d'un ensemble de nombres rationnels (par exemple de l'ensemble des x rationnels tels que $x^2 < 2$) peut fort bien être irrationnelle (dans l'exemple considéré, c'est le nombre $\sqrt{2}$).

Pour les applications à la théorie de l'intégration, nous aurons besoin d'un autre résultat, plus élémentaire.

Théorème 2. Soit A et B deux ensembles non vides de nombres réels, et supposons $x \leqslant y$ pour tout $x \in A$ et tout $y \in B$. Alors A est borné supérieurement, B borné inférieurement, et l'on a $\sup(A) \leqslant \inf(B)$. Pour que $\sup(A) = \inf(B)$, il faut et il suffit que, pour tout entier p, il existe un $x \in A$ et un $y \in B$ tels que l'on ait $y - x \leqslant 10^{-p}$.

Comme B est non vide, il existe des nombres qui appartiennent à B, et qui par suite majorent tout $x \in A$: par suite A est borné supérieurement. A n'étant pas vide, le même raisonnement montre que B est borné inférieurement. Si $y \in B$, on a $x \leqslant y$ pour tout $x \in A$, et donc $\sup(A) \leqslant y$, puisque, par définition, sup A est le *plus petit* nombre qui dépasse tous les $x \in A$. Mais, comme sup(A) est inférieur à tous les $y \in B$, on en conclut, par un argument analogue, que $\sup(A) \leqslant \inf(B)$. Pour tout entier p il y a un $x \in A$ et un $y \in B$ tels que l'on ait :

$$\sup(A) - 10^{-p-1} \leqslant x \leqslant \sup(A) \leqslant \inf(B) \leqslant y \leqslant \inf(B) + 10^{-p-1},$$

d'où :

$$y - x \leqslant \inf(B) - \sup(A) + 2.10^{-p-1} \leqslant \inf(B) - \sup(A) + 10^{-p},$$

et par suite $y - x \leqslant 10^{-p}$ si $\inf(B) = \sup(A)$. Inversement, si l'on peut, pour tout p, trouver un $x \in A$ et un $y \in B$ tels que $y - x \leqslant 10^{-p}$, alors le fait que l'on a $x \leqslant \sup(A) \leqslant \inf(B) \leqslant y$ montre que $\inf(B) - \sup(A) \leqslant 10^{-p}$; cela étant vrai pour tout p, il s'ensuit que $\sup(A) = \inf(B)$.

La notion de borne supérieure d'un ensemble de nombres réels permet de définir celle de *borne supérieure* (ou de « maximum ») *d'une fonction à valeurs réelles*. Soit X un ensemble (par exemple un intervalle dans l'ensemble des nombres réels) et f une fonction définie sur X et à valeurs réelles : f associe donc à chaque $x \in X$ un nombre réel $f(x)$ qui, en général, dépend de x. Notons $f(X)$ l'ensemble des nombres réels y tels qu'il existe un $x \in X$

tel que $y = f(x)$ (« image » de X par f), autrement dit l'ensemble des « valeurs » prises par la fonction $f(x)$ lorsque x « décrit » X. On dit que f est bornée supérieurement (resp. inférieurement) sur X si l'ensemble $f(X)$ est borné supérieurement (resp. inférieurement), c'est-à-dire s'il existe un nombre réel a tel que l'on ait $f(x) \leqslant a$ (resp. $f(x) \geqslant a$) pour tout $x \in X$. Le nombre $\sup(f(X))$ [resp. $\inf(f(X))$] s'appelle alors la borne supérieure (resp. inférieure) ou le maximum (resp. minimum) de la fonction f sur X, et on le désigne par l'assemblage de lettres et de signes que voici :

$$\sup_{x \in X} f(x) \qquad (\text{resp. } \inf_{x \in X} f(x)).$$

On peut donc caractériser le nombre $M = \sup_{x \in X} f(x)$ par les deux propriétés suivantes :

(a) on a $f(x) \leqslant M$ pour tout $x \in X$;

(b) pour tout entier p, il existe un $x \in X$ tel que $M - 10^{-p} < f(x) \leqslant M$.

On fera attention au fait qu'il n'existe pas toujours un $x \in X$ où l'on a *exactement* $f(x) = M$, comme le montre le contre-exemple suivant : on prend pour X l'intervalle $[0, 1[$, ensemble des x réels tels que $0 \leqslant x < 1$, et pour f la fonction $f(x) = x$, dont le graphe est un segment de droite, comme chacun le sait ; on a ici $M = 1$, mais $f(x) < 1$ pour tout $x \in X$.

L'existence d'un nombre réel dont on se donne d'avance des approximations à 10^{-p} près pour tout p peut encore se traduire par le résultat suivant (qui nous sera utile plus loin), habituellement connu sous le nom de « critère de Cauchy », bien que les énoncés qu'on en donne classiquement diffèrent légèrement de celui que l'on trouvera ci-dessous :

Théorème 3. Soit $x_1, x_2, \ldots,$ une suite illimitée de nombres réels. Supposons que, pour tout entier p, il existe un entier q tel que l'on ait $|x_m - x_n| < 10^{-p}$ dès que m et n dépassent q. Alors il existe un nombre réel a tel que, pour tout p, on ait :

$$|x_n - a| \leqslant 10^{-p} \text{ pour tout } n \geqslant q.$$

Nous supposerons (on s'y ramènerait facilement) que tous les nombres x_i sont compris entre 0 et 1. Écrivons le développement décimal de x_i sous la forme :

$$x_i = 0, x_{i_1} x_{i_2} x_{i_3} \ldots,$$

en désignant par x_{i_k} la k-ième décimale de x_i, comprise entre 0 et 9. Comme on a :

$$|x_m - x_n| \leqslant 10^{-p} \text{ pour } m \geqslant q \text{ et } n \geqslant q,$$

il est clair (tout au moins si l'on néglige les difficultés accessoires dues aux développements décimaux impropres) que l'on peut supposer que les $p - 1$ premières décimales de x_m et x_n sont les mêmes dès que m et n dépassent q, autrement dit qu'à partir du rang $n = q$ les $p - 1$ premières décimales de x_n restent fixes. Faisant successivement $p = 2, 3, \ldots,$ on voit donc qu'à partir de l'entier q_2 correspondant à $p = 2$ la première décimale x_{n1} de x_n aura une valeur fixe a_1, puisqu'à partir de l'entier q_3 correspondant à $p = 3$ la seconde décimale x_{n_2} aura en outre une valeur fixe a_2, et ainsi de suite indéfiniment. Considérons alors le nombre réel $a = 0, a_1 a_2 a_3 \ldots,$ dont on vient de construire les décimales de proche en proche ; choisissons un entier p quelconque et considérons l'entier q qui lui correspond d'après l'énoncé du théorème ; pour tout $n \geqslant q$, les $p - 1$ décimales de x_n sont égales à celles de a, par construction même de a. On a donc :

$$|x_n - a| \leqslant 10^{-p+1} \text{ pour tout } n \geqslant q \text{ ;}$$

il reste à montrer qu'on peut en fait, au second membre, remplacer 10^{-p+1} par

10^{-p}. Pour cela choisissons un entier r quelconque ; d'après ce que l'on vient de voir, on a $|x_m - a| \leqslant 10^{-r}$ pour tout m suffisamment grand (remplacer p par $r+1$ dans ce qui précède) ; mais si l'on suppose m et n supérieurs à q on a aussi, par hypothèse, $|x_m - x_n| \leqslant 10^{-p}$; combinant ces deux inégalités, on en conclut que l'on a :

$$|x_n - a| \leqslant 10^{-p} + 10^{-r} \text{ pour tout } n \geqslant q$$

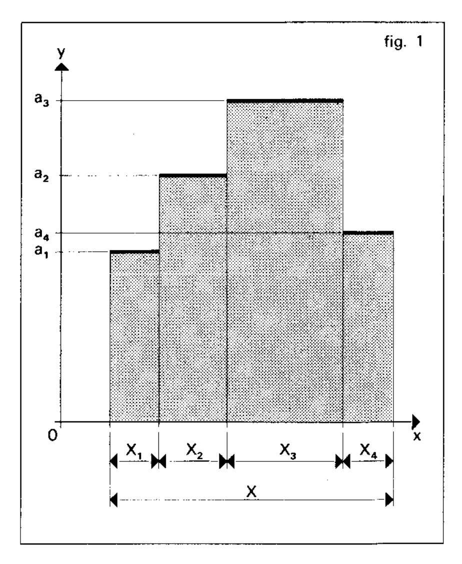
et pour tout entier r. Cela étant valable pour tout r entraîne évidemment l'inégalité cherchée $|x_n - a| \leqslant 10^{-p}$, d'où le théorème.

On énonce habituellement le théorème 3 en termes de suites convergentes et de limites, mais nous ferons en sorte, ici, d'éviter l'usage de ces notions - ce qui est parfaitement possible - afin de faciliter la tâche du lecteur.

2. Intégrale d'une fonction étagée

Considérons, sur un intervalle compact $X = [a, b]$, une fonction $f(x)$ à valeurs réelles ; nous la supposerons même, pour le moment, à valeurs positives. Si l'on trace le graphe de f, on obtient dans le plan une « courbe », l'ensemble des points (x, y) tels que l'on ait $x \in X$ et $y = f(x)$, qui délimite avec l'axe des abscisses et les verticales des points a et b une portion de plan dont on se propose de calculer la surface (dans l'hypothèse où la portion en question serait assez simple pour que l'on puisse attribuer une signification à ce calcul).

Les surfaces les plus simples à calculer sont celles des rectangles. On est ainsi amené à envisager d'abord un cas particulier, celui où l'on peut découper l'intervalle donné X en intervalles partiels $X_1, ..., X_p$ tels que la fonction $f(x)$ prenne, dans chaque X_k, une valeur *constante* a_k ; noter que nous n'excluons aucunement le cas où certains X_k se réduiraient à un point, et que nous n'imposons pas aux X_k d'être compacts — certains X_k peuvent contenir leurs extrémités, d'autres ne pas les contenir. Le graphe de la fonction f se compose alors de segments de droite horizontaux, et on dit que f est une fonction *étagée*. La portion de plan comprise entre le graphe et l'axe des x est alors (fig. 1) réunion de p rectangles ayant pour



bases les intervalles X_k, et pour hauteurs les valeurs a_k correspondantes. Supposant, ce qui est permis, les intervalles X_k deux à deux disjoints (c'est-à-dire sans points communs), nous appellerons alors, par définition, intégrale de f sur l'intervalle X le nombre :

$$(1) \qquad I(f) = a_1 l(X_1) + ... + a_p l(X_p),$$

où $l(X)$ désigne, d'une manière générale, la longueur d'un intervalle X. Le nombre $I(f)$ ainsi obtenu dépend naturellement de

la fonction f considérée, mais non du découpage de l'intervalle X en intervalles partiels X_k sur lesquels f est constante ; si, en effet, l'on découpe, d'une autre façon, X en intervalles deux à deux disjoints $Y_1, ..., Y_q$ sur lesquels f prend les valeurs $b_1, ... b_q$, de sorte que nous devons prouver que :

(2) $$a_1 l(X_1) + ... + a_p l(X_p) = b_1 l(Y_1) + ... + b_q l(Y_q),$$

il est clair que chaque X_k est réunion de ses intersections $X_k \cap Y_h$ avec les divers Y_h ; ces intersections sont des intervalles deux à deux disjoints ; on a donc :

$$l(X_k) = l(X_k \cap Y_1) + ... + l(X_k \cap Y_q),$$

d'où il résulte aussitôt que le premier membre de (2) est la somme de tous les nombres de la forme $a_k l(X_k \cap Y_h)$, le second étant, de même, somme de tous les nombres $b_h l(X_k \cap Y_h)$. Il suffit donc de montrer que l'on a $a_k l(X_k \cap Y_h) = b_h l(X_k \cap Y_h)$ quels que soient k et h. Or la fonction f est égale, sur l'intersection $X_k \cap Y_h$, à la fois à a_k et à b_h ; si l'intersection n'est pas vide, on a donc $a_k = b_h$ et le résultat s'ensuit ; si l'intersection est vide, on a $l(X_k \cap Y_h) = 0$, et le résultat cherché s'écrit $0 = 0$; d'où la relation (2).

Nous utiliserons encore la relation (1) pour définir $I(f)$ lorsque la fonction f n'est plus positive.

Les intégrales des fonctions étagées possèdent des propriétés simples dont nous aurons besoin plus loin. Tout d'abord, si f et g sont deux fonctions étagées sur le même intervalle compact X, alors la fonction $h = f + g$ donnée par $h(x) = f(x) + g(x)$ pour tout $x \in X$ est encore étagée, et l'on a la relation :

(3) $$I(f+g) = I(f) + I(g).$$

En effet, si l'on partage X en intervalles X_i deux à deux disjoints, sur lesquels f prend des valeurs constantes $a_i (1 \leqslant i \leqslant p)$, et en intervalles deux à deux disjoints Y_j, sur lesquels g prend des valeurs constantes $b_j (1 \leqslant j \leqslant q)$, il est clair que X est réunion des pq intervalles $X_i \cap Y_j$, que $f + g$ prend sur $X_i \cap Y_j$ la valeur constante $a_i + b_j$ (d'où le fait que $f + g$ soit étagée), et enfin que $I(f + g)$ est la somme de toutes les expressions $(a_i + b_j)\, l(X_i \cap Y_j)$, c'est-à-dire des expressions $a_i l(X_i \cap Y_j)$, dont la somme est égale à $I(f)$ comme on l'a vu plus haut, et des expressions $b_j l(X_i \cap Y_j)$, dont la somme est égale à $I(g)$ pour des raisons analogues ; d'où la relation (3).

On déduit évidemment de cette relation (3) que $I(f-g) = I(f) - I(g)$; et il est évident que :

(4) $$I(cf) = c.I(f)$$

pour toute constante c, en convenant de définir la fonction étagée $cf = g$ par la condition que $g(x) = c \,.\, f(x)$ pour tout x.

Enfin, si f est une fonction étagée (donc bornée) sur un intervalle compact X, on a l'inégalité :

(5) $$|I(f)| \leqslant l(X). \sup_{x \in X} |f(x)|,$$

où nous utilisons (pour la fonction étagée $|f(x)|$) la notion de borne supérieure définie à la fin du chapitre précédent. En effet, l'inégalité classique $|a + b| \leqslant |a| + |b|$, appliquée à la définition (1) ci-dessus, montre que l'on a :

$$|I(f)| \leqslant |a_1|\, l(X_1) + ... + |a_p|\, l(X_p) ;$$

mais, si l'on pose $M = \sup |f(x)|$, il est clair que M est le plus grand des p nombres $|a_1|, ..., |a_p|$, d'où :

$$|I(f)| \leqslant M.l(X_1) + ... + M.l(X_p) = M[l(X_1) + ... + l(X_p)] = M.l(X),$$

puisque les intervalles X_i sont deux à deux disjoints, et remplissent X tout entier.

3. Intégration des fonctions réglées

Considérons maintenant le cas général ; nous ne supposons plus que la fonction f soit étagée, mais nous supposerons, pour éviter des difficultés secondaires, qu'elle est *bornée*, c'est-à-dire qu'il existe un nombre M tel que l'on ait $-M \leqslant f(x) \leqslant M$ pour tout $x \in X$; pour le moment, supposons aussi $f(x) \geqslant 0$ pour tout x. Soit φ' et φ'' des fonctions étagées telles que l'on ait $0 \leqslant \varphi'(x) \leqslant f(x) \leqslant \varphi''(x)$ pour tout x (il en existe : prendre $\varphi'(x) = 0$ partout, et $\varphi''(x) = M$ partout, par exemple). L'aire limitée par l'axe des x et le graphe de f contient l'aire analogue relative à φ', et est contenue dans l'aire analogue relative à φ'' ; si l'on peut attribuer un sens raisonnable à l'intégrale $I(f)$ de la fonction f, on doit donc avoir la relation :

(6) $$I(\varphi') \leqslant I(f) \leqslant I(\varphi'').$$

On est alors conduit, *que f soit ou non positive*, à introduire deux ensembles E_* et E^* de nombres réels : le premier sera formé des x tels qu'il existe une fonction étagée φ' sur X telle que l'on ait $x = I(\varphi')$ et $\varphi' \leqslant f$ (c'est-à-dire $\varphi'(t) \leqslant f(t)$ pour tout $t \in X$) ; le second sera l'ensemble des nombres réels x pour lesquels on peut trouver sur X une fonction étagée φ'' vérifiant les relations $f \leqslant \varphi''$, et $x = I(\varphi'')$. La relation (6) exprime que le nombre $I(f)$ cherché est supérieur à tout $x \in E_*$, et inférieur à tout $x \in E^*$. Or l'ensemble E^* est borné *inférieurement* et E_* l'est *supérieurement* (les éléments de E_* sont évidemment inférieurs à ceux de E^*, puisque la relation $\varphi' \leqslant \varphi''$, pour des fonctions étagées, implique visiblement l'inégalité $I(\varphi') \leqslant I(\varphi'')$ entre leurs intégrales). Si nous désignons par m la borne supérieure de l'ensemble E_*, et par M la borne inférieure de l'ensemble E^*, la relation (6) exprime que l'intégrale cherchée de f doit être comprise entre m et M ; noter que, comme tout élément de E_* est inférieur à tout élément de E^* comme on vient de le voir, on a aussi $m \leqslant M$; pour que la relation $m \leqslant I(f) \leqslant M$ suffise à déterminer le nombre $I(f)$ cherché, il suffit donc de supposer que l'on a $m = M$. En vertu du théorème 2, les constructions qui précèdent déterminent donc sans ambiguïté $I(f)$ si la condition suivante est remplie : pour tout entier p, il existe sur X des fonctions étagées φ' et φ'' telles que l'on ait d'une part $0 \leqslant \varphi'(x) \leqslant f(x) \leqslant \varphi''(x)$ pour tout $x \in X$, et d'autre part $I(\varphi'') - I(\varphi') \leqslant 10^{-p}$. S'il en est ainsi on dit que la fonction f est *intégrable* (au sens de Riemann) sur l'intervalle X, et l'on pose :

$$\int_a^b f(x)\,dx = I(f).$$

Il est sans doute prudent de déconseiller au lecteur toute tentative d'interprétation *mathématique* de l'assemblage de signes et de lettres figurant au premier membre, et qu'on ne peut expliquer qu'en faisant appel à la psychologie de Leibniz, sujet intéressant, mais qui nous entraînerait trop loin, probablement jusqu'aux philosophes grecs, trop curieux, qui se demandaient comment une étendue finie pourrait bien être obtenue en juxtaposant une infinité d'étendues infinitésimales. La réponse du mathématicien à ce genre de questions est de prier l'interlocuteur de bien vouloir lui fournir, au préalable, une définition *mathématique* des termes qu'il emploie, attendu que c'est la règle de base de toute discussion *mathématique* sérieuse (on trouve même des gens pour prétendre que le respect de cette convention n'est pas sans présenter quelque utilité en dehors des mathématiques) ; et c'est très exactement

parce que personne, en vingt siècles, n'a été capable de lui fournir des définitions précises de ces termes que le mathématicien, en tant que tel, les rejette vers les ténèbres dialectiques de la métaphysique, de la psychologie et de l'histoire humaines, sans chercher à « comprendre » au sens des philosophes. À la limite, la seule chose qui intéresse le mathématicien, en tant que tel, est d'être en mesure de *calculer* des intégrales, voire, diraient beaucoup de gens, d'édifier des procédures mécanisées qui, introduites à l'entrée d'une calculatrice électronique, fourniront automatiquement le résultat cherché avec une marge d'erreur calculable. La définition des intégrales que nous venons d'exposer est éminemment adaptée à ce point de vue « opérationnel » — il suffit de substituer, au calcul de $I(f)$, celui de l'intégrale $I(\varphi)$ d'une fonction étagée φ « suffisamment voisine » de f — comme aussi, bien entendu, au point de vue « absolu » du mathématicien pur, qui désire poursuivre *indéfiniment* la construction du nombre $I(f)$ par approximations de plus en plus étroites. C'est probablement dans ce désir de précision *absolue* que se sont réfugiées les conceptions métaphysiques des Anciens et des classiques, puisqu'on n'imagine pas une machine calculant éternellement...

Indépendamment des problèmes de calcul pratique, qui n'intéressent pas le mathématicien, la notion d'intégrale serait inutilisable si l'on ne connaissait pas « beaucoup » de fonctions intégrables, et si l'on ne disposait pas de procédés mathématiques pour calculer (absolument, c'est-à-dire sans marge d'erreur) les intégrales de beaucoup de fonctions. Nous allons définir ici une catégorie de fonctions intégrables qui, en pratique, permet de couvrir tous les besoins de l'analyse classique.

Pour qu'une fonction f définie et bornée sur un intervalle compact X soit intégrable sur X, il faut et il suffit, comme on l'a vu plus haut, que l'on puisse trouver, pour tout entier $p \geqslant 0$, des fonctions étagées φ' et φ'' sur X vérifiant les relations :

$$(7) \quad \varphi' \leqslant f \leqslant \varphi'', \quad I(\varphi'') - I(\varphi') \leqslant 10^{-p}.$$

La seconde de ces relations s'écrit encore $I(\varphi'' - \varphi') \leqslant 10^{-p}$. Or on a, d'après la relation (5) appliquée à la fonction étagée positive $\varphi'' - \varphi'$, l'inégalité :

$$I(\varphi'' - \varphi') \leqslant l(X). \sup_{x \in X} [\varphi''(x) - \varphi'(x)],$$

en sorte que, pour réaliser la seconde condition (7) ci-dessus, il *suffit* de choisir les fonctions étagées φ' et φ'' de telle sorte que l'on ait :

$$l(X). \sup_{x \in X} [\varphi''(x) - \varphi'(x)] \leqslant 10^{-p};$$

si l'on choisit un entier q tel que $l(X). 10^{-q} \leqslant 10^{-p}$, on est évidemment ramené à construire des fonctions étagées φ' et φ'' vérifiant les conditions suivantes :

$$(8) \quad \varphi' \leqslant f \leqslant \varphi'', \; \varphi''(x) - \varphi'(x) \leqslant 10^{-q}$$

pour tout $x \in X$.

S'il est possible, quel que soit q, de trouver deux telles fonctions φ' et φ'', on dit que la fonction f est *réglée* sur l'intervalle X. Il est bien clair que ces considérations n'ont d'autre but que d'assurer le résultat (*sic*) suivant :

Théorème 4. Soit X un intervalle compact. Toute fonction réglée sur X est intégrable sur X.

Il est à peu près évident que, si f et g sont deux fonctions réglées sur X, il en est de même de leur somme $f + g$, de leur différence $f - g$, et de leur produit $fg = h$, défini par la relation

$h(x) = f(x)g(x)$. On a de plus les relations :

(9) $\quad \mathrm{I}(f+g) = \mathrm{I}(f) + \mathrm{I}(g), \quad \mathrm{I}(cf) = c.\mathrm{I}(f)$

(où c désigne une constante), qui résultent immédiatement des relations analogues (3) et (4) pour les fonctions étagées. Établissons par exemple la première. Pour chaque entier p, on peut par hypothèse trouver des fonctions étagées φ', φ'', ψ', ψ'' vérifiant les conditions suivantes :

$$\varphi' \leqslant f \leqslant \varphi'', \ \mathrm{I}(\varphi'') - \mathrm{I}(\varphi') \leqslant 10^{-p-1},$$
$$\psi' \leqslant g \leqslant \psi'', \ \mathrm{I}(\psi'') - \mathrm{I}(\psi') \leqslant 10^{-p-1},$$

d'où résulte que $\mathrm{I}(f)$ est égal à $\mathrm{I}(\varphi')$ à 10^{-p-1} près, et que $\mathrm{I}(g)$ est égal à $\mathrm{I}(\psi')$ à 10^{-p-1} près aussi. Mais les fonctions étagées $\theta' = \varphi' + \psi'$ et $\theta'' = \varphi'' + \psi''$ vérifient évidemment les relations suivantes :

$$\theta' \leqslant f + g \leqslant \theta'', \ \mathrm{I}(\theta'') - \mathrm{I}(\theta') \leqslant 2.10^{-p-1},$$

la seconde résultant du fait que l'on a :

$$\begin{aligned} &\mathrm{I}(\theta'') - \mathrm{I}(\theta') \\ &= [\mathrm{I}(\varphi'') + \mathrm{I}(\psi'')] - [\mathrm{I}(\varphi') + \mathrm{I}(\psi')] \\ &= [\mathrm{I}(\varphi'') - \mathrm{I}(\varphi')] + [\mathrm{I}(\psi'') - \mathrm{I}(\psi')] \leqslant 2.10^{-p-1}. \end{aligned}$$

Par suite, $\mathrm{I}(f+g)$ est égal à 2.10^{-p-1} près à $\mathrm{I}(\theta') = \mathrm{I}(\varphi') + \mathrm{I}(\psi')$, lui-même égal à 2.10^{-p-1} près à $\mathrm{I}(f) + \mathrm{I}(g)$, de sorte que $\mathrm{I}(f+g)$ et $\mathrm{I}(f) + \mathrm{I}(g)$ sont égaux à 4.10^{-p-1} près, à plus forte raison à 10^{-p} près, et cela pour tout p : d'où la première relation (9).

À partir de la relation (5) établie pour les fonctions étagées on démontrerait, par des raisonnements analogues, que celle-ci est valable plus généralement pour les fonctions réglées. On l'exprime encore, en utilisant la notation de Leibniz, en écrivant que l'on a :

(10) $$\left| \int_a^b f(x)\,dx \right| \leqslant \mathrm{M}\,(b-a)$$

si $|f(x)| \leqslant \mathrm{M}$ pour tout $x \in [a, b]$, résultat connu sous le nom de « premier théorème de la moyenne ». En pratique, il est encore plus utile d'utiliser l'inégalité :

(11) $$\left| \int_a^b f(x)\,dx \right| \leqslant \int_a^b |f(x)|\,dx,$$

qui s'obtient immédiatement sur la formule (1) lorsqu'il s'agit de fonctions étagées, et qu'on étend ensuite facilement aux fonctions réglées par des arguments d'approximation analogues à ceux qu'on a déjà exposé plus haut.

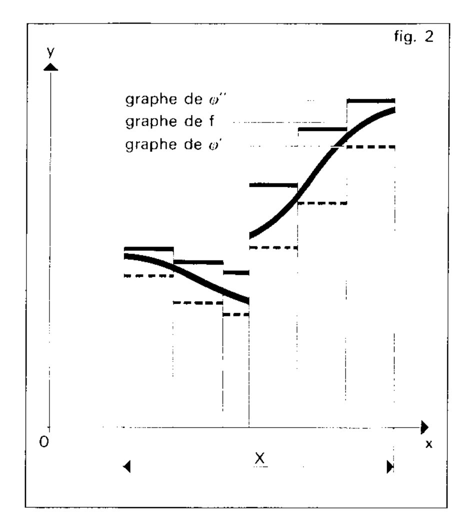
4. Caractérisations des fonctions réglées

Soit f une fonction à valeurs réelles définie sur un intervalle X, et Y un intervalle (ou un ensemble) contenu dans X. Nous dirons que *f est constante à 10^{-p} près sur* Y s'il existe un nombre c tel que l'on ait $|f(x) - c| \leqslant 10^{-p}$ pour tout $x \in \mathrm{Y}$.

Théorème 5. Soit f une fonction définie sur un intervalle compact X. Pour que f soit réglée dans X il faut et il suffit que pour tout entier p, on puisse partager l'intervalle X en un nombre fini d'intervalles partiels $\mathrm{X}_1, ..., \mathrm{X}_n$ tels que f soit constante à 10^{-p} près dans chaque X_i.

C'est presque la définition. Si en effet on peut trouver, pour chaque X_i, une constante c_i telle que $|f(x) - c_i| \leqslant 10^{-p}$ pour tout $x \in \mathrm{X}_i$, on peut immédiatement construire deux fonctions étagées φ' et φ'' vérifiant $\varphi' \leqslant f \leqslant \varphi''$ et $\varphi''(x) - \varphi'(x) \leqslant 2.10^{-p}$ pour tout $x \in \mathrm{X}$, à savoir les fonctions qui dans chaque X_i, sont respectivement égales à $c_i - 10^{-p}$ et à $c_i + 10^{-p}$. Inversement, si f est réglée, il est possible de réaliser les conditions (8) ci-dessus, puis de partager X en intervalles sur chacun desquels φ' et φ'' sont constantes, et sur chacun desquels

f est, par suite, constante à 10^{-q} près (fig. 2).



Dans la pratique, on a besoin de caractériser les fonctions réglées par un critère exprimant que leur comportement « au voisinage de chaque valeur de la variable » est assez simple pour que $f(x)$ n'oscille pas d'une manière incontrôlable lorsque x se rapproche de plus en plus d'une « limite » donnée. Cela va nous obliger à développer quelques considérations sur les *valeurs limites* d'une fonction.

Soit f une fonction réglée sur un intervalle X, et soit a un point de X ; considérons les valeurs de f aux points $x \in X$ tels que $a < x$ (ce qui suppose a distinct de l'extrémité droite de X) ; nous allons montrer que lorsque x se « rapproche de plus en plus » de a, le nombre $f(x)$ « se rapproche de plus en plus » lui aussi d'une certaine valeur b bien définie, que l'on appellera la *valeur limite à droite de f au point a*. Noter que cette propriété est évidente si f est une fonction étagée ; il y a en effet alors un $a' > a$ tel que $f(x)$ soit constant dans l'intervalle $a < x < a'$ (car si l'on décompose X en intervalles partiels sur chacun desquels f est constante, le point a est soit intérieur à l'un de ces intervalles, soit l'extrémité gauche de l'un d'entre eux) ; il est alors clair que si $x > a$ se rapproche de a, $f(x)$, qui ne bouge pas, se rapproche de plus en plus d'une valeur limite (à savoir de sa valeur, constante, dans l'intervalle $a < x < a'$).

Dans le cas général, nous savons qu'il existe, pour chaque entier p, une partition de X en intervalles partiels sur chacun desquels $f(x)$ est constant à 10^{-p} près. Pour chaque entier p, il y a donc dans X un point $a_p > a$ tel que $f(x)$ soit constant à 10^{-p-2} près dans l'intervalle $a < x < a_p$, d'où un nombre réel b_p tel que :

(12) $$a < x < a_p \Rightarrow |f(x) - b_p| \leqslant 10^{-p-2} ;$$

on peut évidemment supposer, si on le désire, que $a_1 \geqslant a_2 \geqslant a_3 \geqslant ...$, en remplaçant au besoin a_p par a_{p-1} pour chaque p tel que $a_p > a_{p-1}$. Cela dit, choisissons un entier p et considérons les nombres b_p, b_{p+1}, b_{p+2}, ... ; si m et n sont deux entiers supérieurs à p, tels par exemple que $p \leqslant m \leqslant n$, considérons un x tel que $a < x < a_n$; on aura aussi les inégalités $a < x < a_m$, d'où, à la fois :

$$|f(x) - b_m| \leqslant 10^{-m-2} \leqslant 10^{-p-2},$$
$$|f(x) - b_n| \leqslant 10^{-n-2} \leqslant 10^{-p-2},$$

ce qui montre évidemment que l'on a :

(13) $$|b_m - b_n| \leqslant 2.10^{-p-2} \leqslant 10^{-p-1},$$

dès que m et n dépassent p. D'après le critère de Cauchy (théorème 3), il existe donc un nombre réel b tel que l'on ait $|b - b_n| \leqslant 10^{-p-1}$ pour tout $n \geqslant p$, et en particulier pour $n = p$. Si $a < x < a_p$, on a alors à la fois $|f(x) - b_p| \leqslant 10^{-p-2}$ et $|b_p - b| \leqslant 10^{-p-1}$, d'où évidemment $|f(x) - b| \leqslant 10^{-p}$.

Cela nous conduit à la notion de limite à droite : on dit qu'une fonction $f(x)$ définie sur un intervalle X possède en un point a de X une *valeur limite à droite* s'il existe un nombre b possédant la propriété que, pour tout p, il existe dans X un nombre $a_p > a$ tel que l'on ait $|f(x)-b| \leqslant 10^{-p}$ pour tout x tel que $a < x < a_p$ (faire attention au fait que l'on n'impose aucune condition pour $x = a$). Cela signifie encore que, pour tout p, il y a un intervalle $a < x < a_p$ d'origine a dans lequel $f(x)$ est égal à b à 10^{-p} près. S'il existe un tel nombre b, il est unique (car si b' satisfait aux mêmes conditions, il existe, pour tout p, des $x > a$ où l'on a à la fois $|f(x)-b| \leqslant 10^{-p-1}$ et $|f(x)-b'| \leqslant 10^{-p-1}$, d'où $|b'-b| \leqslant 10^{-p}$ pour tout p, et finalement $b' = b$). Ce nombre b s'appelle, par définition, la valeur limite à droite de f au point a, et se désigne soit par la notation $f_d(a)$, soit par la notation $f(a+0)$, artifice amusant pour rappeler que l'on considère la limite de l'expression $f(a+h)$ lorsque h tend vers 0 en restant strictement positif. Bien entendu, il ne faut pas, dans la notation précédente, remplacer $a+0$ par a (l'assemblage $a+0$ ne désigne pas, ici, la somme de deux nombres ; il ne constitue qu'un graphisme dépourvu de tout autre sens que celui que nous lui avons attribué conventionnellement). En fait, il peut fort bien arriver que l'on ait $f(a+0) \neq f(a)$, comme le montre l'exemple fort simple de la fonction $f(x)$ définie quel que soit x réel par les formules suivantes :

$$(14) \qquad f(x) = \begin{cases} x & \text{si } x > 1, \\ 0 & \text{si } x = 1, \\ x-4 & \text{si } x < 1. \end{cases}$$

On a ici $f(1+0) = 1$ et $f(1) = 0 \neq 1$.

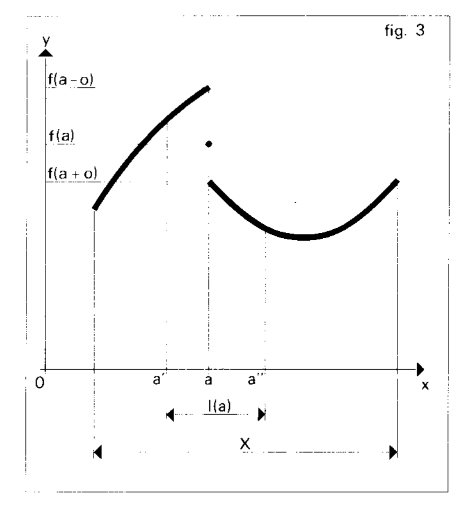
On définirait, bien entendu, de même la notion de valeur limite à gauche de f au point a, notée $f_s(a)$ ou $f(a-0)$: c'est un nombre b (forcément unique s'il existe) tel que, pour tout p, il existe dans X un $a_p < a$ tel que la relation $a_p < x < a$ implique $|f(x)-b| \leqslant 10^{-p}$ (le nombre que nous désignons ici par a_p n'a, bien entendu, aucun rapport avec celui que nous avons noté de la même façon plus haut). Au point où nous en sommes, nous avons évidemment démontré une moitié de l'énoncé fondamental que voici :

Théorème 6. Pour qu'une fonction f définie sur un intervalle compact X soit réglée sur X, il faut et il suffit qu'elle admette en tout point de X des valeurs limites à droite et à gauche.

Il nous reste maintenant à montrer qu'inversement une fonction f qui admet, en chaque point a de X, des valeurs limites à droite et à gauche, est nécessairement réglée sur X. Choisissons pour cela un entier p ; nous devons prouver l'existence d'un nombre *fini* d'intervalles X_1, ..., X_n possédant les propriétés suivantes : X est la réunion de X_1, ..., X_n, et la fonction $f(x)$ est constante à 10^{-p} près sur chacun des X_i.

Mais l'existence de valeurs limites montre que, pour chaque $a \in X$, il existe dans X des nombres a' et a'' tels que l'on ait $a' < a < a''$ et tels que la fonction f soit constante à 10^{-p} près dans chacun des deux intervalles $a' < x < a$ et $a < x < a''$ (si a est l'extrémité gauche de X, on considère seulement a'', et seulement a' si a est l'extrémité droite de X). Désignons alors par I(a) l'intervalle $a' < x < a''$, dans le cas où a n'est pas une extrémité de X, et définissons I(a) de la façon suivante, lorsque a est une extrémité de X : si a est l'extrémité gauche de X, on choisit arbitrairement un $a' < a$

(donc en dehors de X), et on prend pour I(a) l'intervalle $a' < x < a''$, où $a'' \in X$ est choisi comme ci-dessus ; et si a est l'extrémité droite de X, on choisit arbitrairement, en dehors de X, un $a'' > a$, et on prend pour I(a) l'intervalle $a' < x < a''$, où $a' \in X$ est choisi de telle sorte qu'on ait $a' < a$ et que f soit constante à 10^{-p} près dans l'intervalle $a' < x < a$. Les intervalles I(a), pour tous les $a \in X$, jouissent alors des propriétés suivantes (fig. 3) :



(a) chaque I(a) contient a (de sorte que X est contenu dans la réunion des I(a)) ;

(b) chaque I(a) est ouvert, c'est-à-dire de la forme $]a', a''[$, donc défini par des inégalités strictes de la forme $a' < x < a''$;

(c) l'intersection de X et de I(a) peut se décomposer en trois intervalles au plus sur chacun desquels la fonction f est constante à 10^{-p} près ; si par exemple a est distinct des extrémités de X, ces trois intervalles sont $]a', a[$, $[a, a]$ et $]a, a''[$; (noter qu'un intervalle peut fort bien être réduit à un seul point).

En raison de la propriété (c) ci-dessus, il suffirait, pour achever la démonstration du théorème 6, de prouver que l'on peut trouver des points $a_1, ..., a_n \in X$ en nombre fini tels que X soit contenu dans la réunion des intervalles I(a_1), ..., I(a_n), puisque, en décomposant chaque intersection $X \cap X(a_i)$ en, au plus, trois intervalles sur chacun desquels f est constante à 10^{-p} près, on obtiendrait alors une décomposition de X en un nombre fini d'intervalles possédant chacun cette propriété. Tout revient donc, finalement, à établir le résultat suivant, qui peut servir à démontrer beaucoup d'autres théorèmes, et qui se généralise à beaucoup d'autres « espaces topologiques » que les intervalles compacts :

Théorème 7 (Borel-Lebesgue chez les francophones, Heine-Borel à l'étranger). Soit X un intervalle compact. Pour tout $a \in X$, soit I(a) un intervalle ouvert contenant a. Il existe alors des points $a_1, ..., a_n$ de X en nombre fini tels que tout $x \in X$ appartienne à l'un au moins des intervalles I(a_1), ..., I(a_n).

Soit u et v les extrémités gauche et droite de X, et considérons l'ensemble $E \subset X$ des points $x \in X$ qui possèdent la propriété suivante : on peut trouver des points $a_i \in X$ en nombre *fini* tels que tout point de l'intervalle compact $[u, x]$ appartienne à l'un au moins des intervalles I(a_i) ; tout revient à montrer que $v \in E$. On a évidemment $u \in E$, par exemple, parce que l'intervalle $[u, u]$ est contenu dans l'intervalle I(u). L'ensemble E n'étant pas vide, et étant borné supérieurement (on a $x \leqslant v$ pour tout $x \in E$) admet donc (théorème 1) une borne supérieure $a \leqslant v$; nous allons montrer que $a \in E$, puis que $a = v$, ce qui terminera la démonstration.

Considérons en effet l'intervalle ouvert $I(a) =]a', a''[$ avec $a' < a < a''$. Comme $a = \sup(E)$, il existe un $x \in E$ tel que

$a' < x < a$. Comme $x \in$ E, l'intervalle $[u, x]$ est contenu dans la réunion d'un nombre fini d'intervalles $\mathrm{I}(a_1), ..., \mathrm{I}(a_p)$; posant $a_{p+1} = a$, il est alors clair que l'intervalle $[u, a]$, réunion de $[u, x]$ et de $[x, a]$, est contenu dans la réunion de $\mathrm{I}(a_1), ..., \mathrm{I}(a_{p+1})$. Cela montre que $a \in$ E.

Montrons enfin que $a = v$. Supposons en effet $a < v$. Posant $\mathrm{I}(a) =]a', a''[$, comme ci-dessus, il existe évidemment des $x \in$ X tels que $a < x < a''$; pour un tel x, l'intervalle $[u, x]$, réunion de $[u, a]$ et de $[a, x]$, est contenu dans la réunion de $[u, a]$ et de $\mathrm{I}(a)$; comme l'on peut « recouvrir » $[u, a]$ à l'aide d'un nombre fini d'intervalles $\mathrm{I}(a_i)$ comme on vient de le démontrer, il en est donc de même de $[u, x]$. Par suite, $x \in$ E ; ce qui est absurde puisque x est strictement supérieur au nombre $a = \sup$ (E). On a donc bien $a = v$, ce qui démontre le théorème 7, ainsi par conséquent que le théorème 6.

Les exemples les plus simples de fonctions réglées sont les *fonctions continues*. On dit qu'une fonction f est continue en un point a si elle admet en ce point des valeurs limites à droite et à gauche et si de plus on a $f(a-0) = f(a) = f(a+0)$. Il revient évidemment au même de dire qu'une fonction f, définie sur un intervalle X (compact ou non), est continue en a si, pour tout p, il existe un intervalle ouvert $\mathrm{I}(a)$ contenant a et tel que f soit constante à 10^{-p} près dans $\mathrm{X} \cap \mathrm{I}(a)$. Lorsque f est continue en tout point de X, on dit que f est continue sur X. Le théorème 4 montre alors que, *sur un intervalle compact, toute fonction continue est intégrable*, résultat classique obtenu par Darboux en 1875, et dont on peut aujourd'hui donner, à partir de rien (la notion expérimentale de nombre réel conçu comme développement décimal illimité), une démonstration parfaitement rigoureuse et complète en quelques pages.

Parmi les fonctions réglées figurent aussi les fonctions *monotones*, c'est-à-dire les fonctions croissantes et les fonctions décroissantes ; une fonction f, sur un intervalle X, est dite *croissante* dans X si, quels que soient $x', x'' \in$ X, la relation $x' \leqslant x''$ implique $f(x') \leqslant f(x'')$ – la valeur de $f(x)$ augmente lorsque x augmente – et *décroissante* si, au contraire, la relation $x' \leqslant x''$ implique $f(x') \geqslant f(x'')$; ne pas se fier, pour ces notions, aux définitions bizarres et compliquées, bien que prétendument « intuitives », que l'on trouve encore dans de nombreux manuels. Pour montrer qu'une fonction f croissante, par exemple, est réglée, on considère un point quelconque $a \in$ X et l'ensemble E des valeurs $f(x)$ prises par f aux points $x \in$ X tels que $x > a$; comme f est croissante, on a toujours $f(x) \geqslant f(a)$ pour un tel x, et par suite E est borné inférieurement, donc admet une borne inférieure $b = \inf$ (E), dont nous allons montrer que c'est précisément la limite à droite $f(a+0)$, dont l'existence sera ainsi établie. Soit, en effet, p un entier quelconque ; il y a dans E un nombre qui est égal à b à 10^{-p} près, donc un $a' \in$ X vérifiant les relations $a' > a$ et $b \leqslant f(a') \leqslant b + 10^{-p}$. Pour $a < x < a'$, on a alors d'une part $b \leqslant f(x)$ puisque $f(x) \in$ E, et d'autre part aussi $f(x) \leqslant f(a')$, puisque f est croissante ; il s'ensuit que l'on a $b \leqslant f(x) \leqslant b + 10^{-p}$, dès que $a < x < a'$, ce qui prouve très exactement que f admet, au point a, une valeur limite à droite égale à b. On établirait de même l'existence de valeurs limites à gauche.

5. Intégration et dérivation

Soit X un intervalle quelconque, et f une fonction réglée dans X, c'est-à-dire qui admet des limites à gauche et à droite en

chaque point de X. Choisissons une fois pour toutes un point a de X. Pour tout $t \in X$, la fonction f est réglée et donc intégrable dans l'intervalle compact d'extrémités a et t. Nous nous proposons d'étudier la fonction F définie sur X par les formules suivantes :

$$(15) \qquad F(t) = \int_a^t f(x)\,dx \quad \text{si } a \leqslant t,$$

$$F(t) = -\int_t^a f(x)\,dx \quad \text{si } t \leqslant a.$$

Notons qu'en convenant de définir :

$$(16) \quad \int_a^b f(x)\,dx = -\int_b^a f(x)\,dx, \quad \text{lorsque } a \geqslant b$$

on a :

$$F(t) = \int_a^t f(x)\,dx \quad \text{pour tout } t \in X\,;$$

nous dirons que F est la *primitive de f au point a*. On a évidemment $F(a) = 0$.

Les deux outils essentiels pour l'étude de F sont d'une part l'inégalité (10) démontrée plus haut, et d'autre part la relation :

$$(17) \qquad \int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

valable quels que soient $a, b, c \in X$, et dont l'analogie avec la célèbre « relation de Chasles » :

$$\overline{AB} = \overline{AC} + \overline{CB}$$

de la théorie des segments de droite orientés est évidente. La démonstration de (17) consiste d'abord à utiliser la relation (16) et des calculs triviaux pour ramener la démonstration de (17) au cas où l'on a $a \leqslant c \leqslant b$; on vérifie alors (17) immédiatement lorsque f est une fonction étagée en utilisant la définition (1) de l'intégrale, puis on passe de là au cas général en appliquant (17) à des fonctions étagées « très voisines » de f, et en utilisant des arguments d'approximation analogues à ceux qui nous ont servi pour établir par exemple le théorème 4. Intuitivement, la relation (17) exprime que l'aire comprise entre le graphe de f, l'axe des x, et les verticales a et b, est somme de l'aire analogue comprise entre les verticales a et c, et de l'aire analogue comprise entre les verticales c et b, ce qui est « évident géométriquement », c'est-à-dire aussi longtemps qu'on n'exige pas de véritable démonstration...

Cela fait, revenons à la fonction $F(t)$. Pour un $t \in X$ distinct de l'extrémité droite de X, on a $t + h \in X$ pour tout nombre $h > 0$ suffisamment petit. Comme f admet en t une limite à droite $f(t+0)$, il y a, pour tout entier p, un nombre $t_p > t$ dans X tel que la relation $t < t + h < t_p$ implique $|f(t+h) - f(t+0)| \leqslant 10^{-p}$. Posons $b = f(t+0)$ et considérons, pour un tel h, l'intégrale :

$$(18) \qquad \int_t^{t+h} [f(x) - b]\,dx = \int_t^{t+h} f(x)\,dx - \int_t^{t+h} b.dx = F(t+h) - F(t) - bh,$$

puisque :

$$F(t+h) - F(t) = \int_a^{t+h} f(x)dx - \int_a^t f(x)dx = \int_t^a f(x)\,dx + \int_a^{t+h} f(x)\,dx = \int_t^{t+h} f(x)\,dx.$$

Comme $|f(x) - b| \leqslant 10^{-p}$ pour $t < x < t + h$, et comme les valeurs de $f(x) - b$, aux extrémités, n'ont évidemment aucune influence sur le calcul de l'intégrale, le premier théorème de la moyenne (10) montre que l'on a :

$$\left| \int_t^{t+h} [f(x) - b]\,dx \right| \leqslant 10^{-p}.h\,;$$

de là et de (18) résulte donc que l'on a :

$$(19) \quad |F(t+h) - F(t) - bh| \leqslant 10^{-p}.h$$

dès que $t < t + h < t_p$, ou, puisque $h > 0$, que :

$$(20) \quad \left|\frac{F(t+h) - F(t)}{h} - b\right| \leqslant 10^{-p}$$

dès que $t < t + h < t_p$.

On est ainsi conduit à dire qu'une fonction F admet en un point t une *dérivée à droite* égale à b si, pour tout entier p, il existe un nombre $t_p > t$ tel que l'on ait la relation (20), ou, si l'on préfère (poser $t_p = t + h_p$), un nombre $h_p > 0$ tel que la relation :

$$(21) \quad 0 < h < h_p \text{ implique : } \left|\frac{F(t+h) - F(t)}{h} - b\right| \leqslant 10^{-p}$$

le rapport $[F(t+h) - F(t)]/h$ représente, géométriquement, la pente de la droite joignant les points $(t, F(t))$ et $(t + h, F(t+h))$ du graphe de F. On désigne habituellement la dérivée à droite au point t, si elle existe, par $F'_d(t)$.

On définit bien entendu de même la notion de *dérivée à gauche* au point t : c'est un nombre c possédant la propriété que, pour tout p, il existe un nombre $h_p < 0$ (aucun rapport avec ce que nous avons désigné ci-dessus par h_p) tel que :

$$(22) \quad h_p < h < 0 \Rightarrow \left|\frac{F(t+h) - F(t)}{h} - c\right| \leqslant 10^{-p}$$

Le raisonnement qui, pour la primitive F de f, nous a conduit à la relation (19) avec $b = f(t+0)$ conduirait, moyennant des modifications triviales, à la relation (22) avec cette fois $c = f(t-0)$. En résumé :

Théorème 8. Soit f une fonction réglée sur un intervalle X, et

$$F(t) = \int_a^t f(x)\,dx$$

la primitive de f en un point a de X. Alors la fonction F admet en chaque point $t \in X$ des dérivées à droite et à gauche données par :

$$(23) \quad F'_d(t) = f(t+0),\ F'_g(t) = f(t-0).$$

Le cas le plus fréquent est celui d'une fonction f partout *continue* : alors $F'_d(t) = F'_g(t)$ pour tout $t \in X$, ce qui exprime que la fonction F admet en chaque point $t \in X$ une *dérivée* $F'(t) = F'_g(t) = F'_d(t)$, donnée, puisque $f(t-0) = f(t+0) = f(t)$, par :

$$(24) \qquad F'(t) = f(t).$$

Nous allons maintenant montrer que cette relation suffit pour caractériser presque entièrement la fonction F, de sorte que *le calcul des intégrales portant sur la fonction f reviendra à la détermination d'une solution* F *de l'équation (24)* ; c'est là le cœur même du « calcul infinitésimal » tel que le concevaient les analystes du XVIIe siècle.

6. Détermination d'une fonction par sa dérivée

Soit F et G, deux fonctions admettant en un point t des dérivées $F'(t)$ et $G'(t)$; on voit facilement qu'alors les fonctions $F + G$, $F - G$ et FG admettent aussi, au point t, des dérivées données par des formules simples, et égales respectivement à :

$$F'(t) + G'(t), \quad F'(t) - G'(t),$$
$$F'(t)G(t) + F(t)G'(t).$$

La meilleure façon d'établir ces résultats est d'utiliser la notation de Landau et d'écrire, ce qui est clair d'après (19),

qu'une fonction F admet au point t une dérivée égale à b si et seulement si l'on a la relation :

(25) $$F(t + h) = F(t) + bh + o(h)$$

lorsque h tend vers 0 ; écrivant de même que :

$$G(t + h) = G(t) + ch + o(h)$$

lorsque h tend vers 0, avec $c = G'(t)$, on en déduit par addition que :

$$F(t + h) + G(t + h) = F(t) + G(t) + (b + c)h + o(h),$$

d'où l'existence et le calcul de la dérivée de $F + G$...

Cela dit, et en nous bornant aux fonctions dérivables pour éviter des complications secondaires (mais les résultats que nous allons établir seraient encore valables, moyennant des modifications triviales, pour des fonctions admettant partout des dérivées à droite et à gauche), considérons deux solutions F_1 et F_2 de l'équation (24) ; la fonction $F = F_1 - F_2$ admet alors dans l'intervalle X considéré une dérivée :

$$F'(t) = F'_1(t) - F'_2(t) = f(t) - f(t) = 0.$$

On est alors amené à établir le résultat suivant :

Théorème 9. Soit F une fonction définie dans un intervalle X et admettant, en tout point de X, une dérivée égale à 0. Alors la fonction F est constante dans X.

Si l'on admet ce théorème, on voit que deux solutions quelconques de (24) diffèrent entre elles d'une simple *constante*. Si donc l'on connaît une solution F de (24), et si l'on choisit un point a de X, on aura une relation de la forme :

(26) $$\int_a^t f(x)\,dx = F(t) + c,$$

où c est une constante indépendante de t. En particulier, pour $t = a$, on obtient la relation :

$$0 = F(a) + c,$$

d'où : $c = -F(a)$ et $\int_a^t f(x)\,dx = F(t) - F(a)$.

Autrement dit :

Théorème 10. Soit f une fonction continue dans un intervalle X et F une fonction dérivable dans X telle que $F'(t) = f(t)$ pour tout $t \in X$. On a alors :

(27) $$\int_a^b f(x)\,dx = F(b) - F(a),$$

quels que soient a, $b \in X$.

C'est ce qu'on appelle fréquemment le *théorème fondamental du calcul infinitésimal* puisqu'il montre l'équivalence entre les deux problèmes suivants : calculer

$$\int_a^b f(x)\,dx$$

pour toutes les valeurs possibles de a et b ; trouver une fonction F telle que $F'(t) = f(t)$ quel que soit t. Si, par exemple, l'on sait que la dérivée de la fonction x^{15} est $15x^{14}$, de sorte qu'on a $F'(t) = f(t)$ si $f(t) = x^{14}$ et $F(t) = x^{15}/15$, on peut immédiatement écrire :

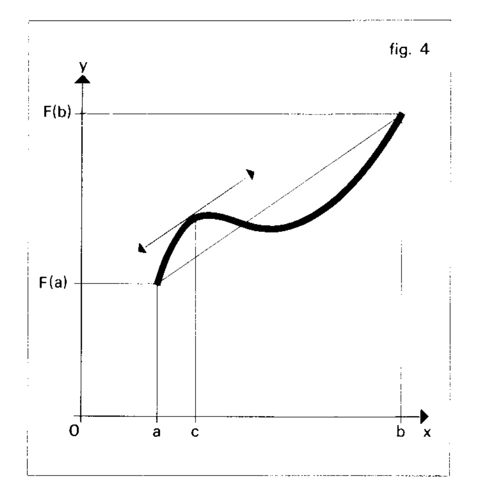
$$\int_a^b x^{14}\,dx = \frac{b^{15} - a^{15}}{15},$$

quels que soient a et b, sans autre calcul.

7. Théorème des accroissements finis

Il nous faut maintenant démontrer le théorème 9, c'est-à-dire prouver que l'on a $F(a) = F(b)$, quels que soient a et b dans X. L'idée de la démonstration est d'une extrême simplicité, et fort ingénieuse. Elle consiste à observer que, si l'on contemple

le graphe F dans l'intervalle $]a, b[$, on peut trouver un point c de cet intervalle où la tangente au graphe de F (fig. 4) est



parallèle à la « corde » joignant les points $(a, F(a))$ et $(b, F(b))$ du graphe ; si $F(a) \neq F(b)$, cette corde n'est pas horizontale, la tangente en c non plus, et l'on a par suite $F'(c) \neq 0$, contrairement à l'hypothèse !

Mais ce raisonnement purement géométrique doit être rendu rigoureux grâce à une démonstration effective de l'existence du point c en question. Noter que la pente de la corde joignant les points $(a, F(a))$ et $(b, F(b))$ est égale à :

$$\frac{F(b) - F(a)}{b - a}.$$

Nous sommes donc ramenés, pour établir le théorème 9, à établir le résultat bien plus utile encore que voici :

Théorème 11 (formule des accroissements finis). Soit F une fonction dérivable dans un intervalle compact $[a, b]$. Il existe un point $c \in]a, b[$ où l'on a :

$$(28) \qquad F'(c) = \frac{F(b) - F(a)}{b - a}.$$

(Par une fonction dérivable sur $[a, b]$ nous entendons une fonction qui admet une dérivée en tout point x tel que $a < x < b$, ainsi qu'une dérivée à droite en a et une dérivée à gauche en b. On démontre souvent le théorème 11 sans supposer l'existence de ces dérivées en a et b.)

Notons d'abord qu'*il suffit d'établir le théorème 11 pour les fonctions* F *telles que* $F(a) = F(b)$. Supposons-le en effet établi moyennant cette hypothèse supplémentaire, et substituons à la fonction F donnée la fonction :

$$G(t) = F(t) - \frac{F(b) - F(a)}{b - a}(t - a);$$

on a évidemment $G(a) = F(a)$, et $G(b) = F(b) - [F(b) - F(a)] = F(a) = G(a)$. De plus, il est clair que G admet partout dans X une dérivée donnée par :

$$(29) \qquad G'(t) = F'(t) - \frac{F(b) - F(a)}{b - a},$$

puisque la dérivée d'une fonction linéaire $ct + d$ est égale à c. Établi pour G, le théorème 11 exprime l'existence d'un nombre :

$$c \in]a, b[\text{ où } G'(c) = \frac{G(b) - G(a)}{b - a} = 0,$$

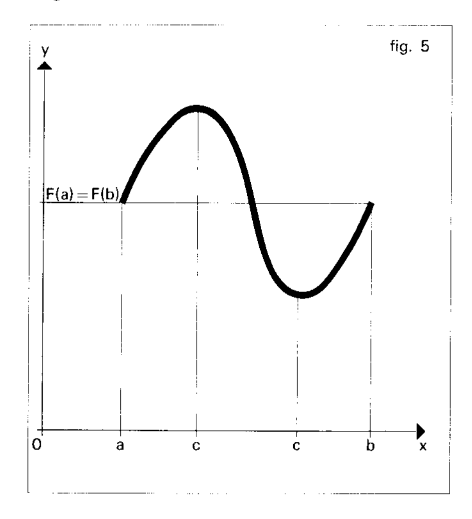
ce qui, d'après (29), fournit aussitôt la relation (28) pour la fonction F.

Nous pouvons donc bien supposer $F(a) = F(b) = 0$; le théorème 11 s'énonce alors comme suit :

Théorème 11 bis. Soit F une fonction dérivable dans un intervalle compact $[a, b]$ et telle que $F(a) = F(b)$. Il existe un point $c \in]a, b[$ où l'on a $F'(c) = 0$.

Ce résultat est connu sous le nom de « théorème de Rolle », académicien français de la fin du XVIIe siècle, et resté célèbre pour avoir eu le premier, du moins le suppose-t-on, l'idée géométrique d'une

démonstration du théorème 11 *bis*. Cette idée, identique à celle que nous avons exposée au début de ce chapitre, consiste à remarquer que le théorème 11 *bis* exprime l'existence d'un point du graphe de F où la tangente à celui-ci est *horizontale* ; et la meilleure façon de trouver un tel point (c'est du moins l'impression que l'on retire d'une réflexion géométrique simple) consiste à le chercher parmi les points où la fonction F est *maximum* ou *minimum* (fig. 5).



Supposons en effet trouvé un point c vérifiant les conditions suivantes ; on a :

(30) $\quad a < c < b$ et $F(x) \leqslant F(c)$

pour tout $x \in [a, b]$. Le rapport :

$$\frac{F(c+h) - F(c)}{h},$$

défini pour tout $h \neq 0$ assez petit (parce qu'on suppose c distinct des extrémités a et b de l'intervalle considéré) est alors positif ou nul pour $h < 0$ (quotient de deux nombres négatifs), et négatif ou nul pour $h > 0$. Mais si $|h|$ est suffisamment petit, ce rapport est égal, à 10^{-p} près, à $F'(c)$; on en conclut que $F'(c)$ doit être à la fois positif et négatif, ce qui prouve que $F'(c) = 0$. On parviendrait évidemment à la même conclusion en supposant :

(31) $\quad a < c < b$ et $F(x) \geqslant F(c)$

pour tout $x \in [a, b]$.

On voit ainsi qu'en définitive les théorèmes 9, 10 et 11 reposent sur l'existence d'un point c vérifiant soit les conditions (30), soit les conditions (31).

8. Théorème du maximum

Comme nous allons le voir, il suffit pour cela d'établir le résultat suivant :

Théorème 12. Soit F une fonction définie et continue sur un intervalle compact X. Il existe un $c' \in X$ tel que l'on ait $F(x) \leqslant F(c')$ pour tout $x \in X$, et un $c'' \in X$ tel que l'on ait $F(c'') \leqslant F(x)$ pour tout $x \in X$.

Avant d'établir le théorème 12, montrons comment il implique le théorème 11 *bis*. Tout d'abord la fonction F du théorème 11 *bis*, étant dérivable en tout point de X, est continue. En effet, pour tout $t \in X$, il existe, d'après les relations (21) et (22), que l'on écrira pour $p = 0$, des nombres h' et h'' tels que l'on ait $h' < 0 < h''$ et tels que :

(32) $\quad h' < h < h'' \Rightarrow$
$$|F(t+h) - F(t) - F'(t)h| \leqslant |h|$$

(le cas où $h = 0$ est exclu de (21) et (22), mais se traite directement par vérification triviale). Posant $M = 1 + |F'(t)|$, on en déduit que l'on a $|F(t+h) - F(t)| \leqslant M|h|$ pour $h' < h < h''$. Choisissons un entier n tel que $M \leqslant 10^n$; alors $M|h| \leqslant 10^{-p}$ pourvu que $|h| \leqslant 10^{-p-n}$. Les t' tels que le nombre $h = t' - t$ vérifie à la fois $h' < h < h''$ et $|h| \leqslant 10^{-p-n}$ forment un inter-

valle ouvert I(t) contenant le point t, et le raisonnement précédent montre que, pour $t' \in X \cap I(t)$, on a $|F(t') - F(t)| \leqslant 10^{-p}$, autrement dit que F est constante à 10^{-p} près dans $X \cap I(t)$: d'où la continuité de F en t (cf. fin du chap. 5).

Cela dit, le théorème 12 s'applique à F, d'où des points c' et c'' dans X tels que l'on ait $F(c'') \leqslant F(x) \leqslant F(c')$ pour tout $x \in X$. Si l'on a $a < c' < b$, la condition (30) est réalisée ; si l'on a $a < c'' < b$, la condition (31) l'est, et dans chacun de ces deux cas le théorème 11 *bis* est démontré, comme on l'a vu. Il reste à examiner le cas où c' et c'' sont situés aux extrémités de X. Mais comme $F(a) = F(b)$, on a alors $F(c') = F(c'') = F(a) = F(b)$, donc $F(x) = F(a) = F(b)$ pour tout $x \in X$, et la fonction F est constante, d'où $F'(t) = 0$, quel que soit $t \in X$, de sorte que le théorème 11 *bis* est trivialement vrai dans ce cas aussi.

Passons maintenant à la *démonstration du théorème 12*. Tout d'abord la fonction, étant continue sur l'intervalle compact X, est réglée sur X d'après le théorème 6, comme nous l'avons déjà observé au chapitre 5 ; elle est donc *bornée* sur X : choisir sur X une fonction étagée φ telle que l'on ait par exemple $|F(x) - \varphi(x)| \leqslant 1$ pour tout $x \in X$, désigner par u et v la plus petite et la plus grande des valeurs (en nombre fini) prises par la fonction φ sur X, et observer qu'on a alors $u - 1 \leqslant F(x) \leqslant v + 1$ pour tout $x \in X$. L'ensemble F(X) des valeurs prises par la fonction F sur l'intervalle X est donc borné et admet par suite une borne supérieure M et une borne inférieure m (cf. chap. 1) ; toute la question est de prouver l'existence de nombres c' et $c'' \in X$ tels que $m = F(c'')$ et $M = F(c')$ (ce qui, nous l'avons vu au chapitre 1, pourrait fort bien se révéler impossible si l'intervalle X n'était pas supposé être *compact*), et nous nous bornerons à prouver l'existence de c', celle de c'' se démontrant de la même façon (ou, mieux encore, se déduisant de l'existence de c' puisque c'' joue pour la fonction $-F$ le même rôle que c' pour la fonction F).

Or nous savons que, pour tout p, il existe des $x \in X$ où l'on a $M - 10^{-p} \leqslant F(x) \leqslant M$. Soit A_p l'ensemble de ces $x \in X$; tout revient à prouver qu'il existe un point c commun à tous les A_p, car si l'on a :

$$M - 10^{-p} \leqslant F(c) \leqslant M$$

quel que soit p, on aura évidemment aussi $F(c) = M$. Nous allons maintenant raisonner par l'absurde, en supposant que les ensembles A_p n'ont aucun point commun et en déduisant de là une contradiction.

Si les A_p n'ont aucun point commun, alors, pour tout $x \in X$, il existe un entier p tel que $x \notin A_p$, c'est-à-dire tel que l'on ait $F(x) < M - 10^{-p}$. Comme F est *continue* en tout point de X, il existe alors un intervalle ouvert I(x) contenant x, et tel que F soit constante à 10^{-p-1} près dans $X \cap I(x)$; pour tout $x' \in X \cap I(x)$ on a donc alors $F(x) \leqslant M - 10^{-p} + 10^{-p-1} < M - 10^{-p-1}$ puisque l'on a :

$$-10^{-p} + 10^{-p-1} = \frac{-1}{10^p} + \frac{1}{10^{p+1}} = \frac{-10+1}{10^{p+1}} = \frac{-9}{10^{p+1}} < \frac{-1}{10^{p+1}},$$

comme on le voit en observant que $-9 < -1$. Désignant par q l'entier $p + 1$ on voit donc que, pour tout $x \in X$, il existe un intervalle ouvert I(x) contenant x et un entier q tels que l'on ait :

$$(33) \qquad F(x') < M - 10^{-q}$$

pour tout $x' \in I(x) \cap X$.

Appliquons alors le théorème 7 ; on trouve des points $x_1, ..., x_n$ de X, en nom-

bre fini, et des entiers $q_1, ..., q_n$, tels que les conditions suivantes soient remplies ; tout $x \in X$ appartient à l'un au moins des intervalles $I(x_1), ..., I(x_n)$, et d'autre part on a, d'après (33) appliqué à chaque x_k,

$$(34) \qquad F(x) < M - 10^{-q_k}$$

pour tout $x \in I(x_k) \cap X$.

Soit q le plus grand des n entiers $q_1, ..., q_n$; on a évidemment $10^{q_k} \leqslant 10^q$, et donc :

$$10^{-q_k} \geqslant 10^{-q}, \text{ d'où } M - 10^{-q_k} \leqslant M - 10^{-q},$$

pour tout k tel que $1 \leqslant k \leqslant n$; par suite, (34) montre que $F(x) < M - 10^{-q}$ pour tout $x \in X$ qui appartient à l'un au moins des $I(x_k)$, c'est-à-dire pour tout $x \in X$ sans exception. Mais on aboutit ici à une contradiction puisque, par définition de la borne supérieure M de la fonction F, il existe certainement des $x \in X$ où l'on a $F(x) \geqslant M - 10^{-q}$. Le théorème 12 est donc démontré, et avec lui les théorèmes 11, 10 et 9.

9. Formule de Taylor

Nous allons maintenant établir le dernier « grand » résultat de l'analyse infinitésimale, à savoir la formule de Taylor, qui permet, au voisinage d'un point, de remplacer une fonction « suffisamment régulière » par un polynôme qui lui est « approximativement » égal.

Soit f une fonction définie dans un intervalle ouvert X. Nous dirons que f est *de classe* C^1 dans X si elle admet une dérivée $f'(x)$ en tout point de X, et si celle-ci est fonction continue de x. Si f' est elle-même de classe C^1, on dit que f est de classe C^2 ; on peut alors attribuer à f une *dérivée seconde* continue $f'' = (f')'$. En poursuivant ainsi de proche en proche on définit de façon évidente les dérivées successives, et la notion de *fonction de classe* C^p, c'est-à-dire admettant dans X des dérivées continues jusqu'à l'ordre p inclusivement.

Supposons f de classe C^1 dans X, et appliquons le théorème 10 à la fonction continue f' ; on trouve que :

$$(35) \qquad f(b) - f(a) = \int_a^b f'(x)\,dx,$$

quels que soient $a, b \in X$.

Soit maintenant u et v deux fonctions de classe C^1 dans X ; la fonction $w = uv$ l'est aussi et sa dérivée est donnée par la formule $w' = u'v + uv'$, que nous avons déjà indiquée après l'énoncé du théorème 11. On en déduit que :

$$\begin{aligned} w(b) - w(a) &= \int_a^b [u'(x)\,v(x) + u(x)\,v'(x)]\,dx \\ &= \int_a^b u'(x)\,v(x)\,dx + \int_a^b u(x)\,v'(x)\,dx, \end{aligned}$$

d'où la *formule d'intégration par parties* :

$$(36) \quad \int_a^b u'(x)\,v(x)\,dx = u(b)\,v(b) - u(a)\,v(a) - \int_a^b u(x)\,v'(x)\,dx,$$

valable pour u et v de classe C^1 dans X, très commode pour le calcul pratique des intégrales, mais dont l'intérêt est ailleurs lorsqu'on s'occupe de mathématiques. Nous allons obtenir la formule de Taylor en combinant les formules (35) et (36).

Pour cela supposons, dans (35), que f soit de classe C^2 et donc f' de classe C^1. Appliquons la formule (36) en choisissant $u(x) = x - b$, d'où $u'(x) = 1$, et $v(x) = f'(x)$, d'où $v'(x) = f''(x)$; il vient $u(b) = 0$, $u(a) = a - b$, $v(b) = f'(b)$, $v(a) = f'(a)$, d'où :

$$\begin{aligned} f(b) - f(a) &= \int_a^b f'(x)\,dx \\ &= f'(a)\,(b - a) - \int_a^b f''(x)\,(x - b)\,dx. \end{aligned}$$

Supposons maintenant f de classe C^3, donc f'' de classe C^1 ; on peut calculer la dernière intégrale en faisant $u(x) = f''(x)$ et $v(x) = (x-b)^2/2$ dans la formule (36), puisque alors $v'(x) = x - a$; il vient donc, puisque $v(b) = 0$:

$$\begin{aligned} f(b)-f(a) &= f'(a)(b-a) - \{-f''(a)(a-b)^2/2 \\ &\quad - \int_a^b f'''(x)\frac{(x-b)^2}{2}dx\} \\ &= f'(a)(b-a) + f''(b-a)^2/2 \\ &\quad + \int_a^b f'''(x)\frac{(x-b)^2}{2}dx. \end{aligned}$$

Si f est de classe C^4, on peut calculer la dernière intégrale en prenant, dans la formule d'intégration par parties, $u(x) = f'''(x)$ et $v(x) = (x-b)^3/2.3$; comme $v(b) = 0$, il vient alors manifestement :

$$\begin{aligned} f(b)-f(a) &= f'(a)(b-a) + f''(a)(b-a)^2/2 \\ &+ f'''(a)(b-a)^3/2.3 - \int_a^b f''''(x)\frac{(x-b)^3}{2.3}dx. \end{aligned}$$

Il est clair que le raisonnement se poursuit aussi longtemps que les dérivées existent et sont continues. Autrement dit, si f est de classe C^{p+1}, on a la relation :

$$(37)\quad \begin{aligned} f(b)-f(a) &= f'(a)(b-a) + f''(a)(b-a)^2/2 + \ldots \\ &+ f^{(p)}(a)(b-a)^p/p\,! + R_p, \end{aligned}$$

où l'on a posé $p\,! = 1.2....p$ (produit des p premiers nombres entiers), et où l'expression R_p est donnée par la relation :

$$(38)\quad \begin{aligned} R_p &= (-1)^p \int_a^b f^{(p+1)}(x)\frac{(x-b)^p}{p\,!}dx \\ &= \int_a^b f^{(p+1)}(x)\frac{(b-x)^p}{p\,!}dx. \end{aligned}$$

La *formule de Taylor avec reste intégral* est la relation :

$$(39)\quad \begin{aligned} f(b) &= f(a) + f'(a)(b-a) + f''(a)(b-a)^2/2! + \ldots \\ &+ f^{(p)}(a)(b-a)^p/p\,! + R_p, \end{aligned}$$

qui se déduit immédiatement de (37). Si l'on fixe le point a en remplaçant b par un point variable $t \in X$, on obtient la relation :

$$(40)\quad \begin{aligned} f(t) &= f(a) + f'(a)(t-a) + f''(a)(t-a)^2/2! + \ldots \\ &+ f^{(p)}(a)(t-a)^p/p! + R_p(t), \end{aligned}$$

qui exprime f comme somme d'un *polynôme* en t et d'un « reste » :

$$(41)\quad R_p(t) = \int_a^t f^{(p+1)}(x)\frac{(t-x)^p}{p\,!}dx.$$

Noter que le polynôme en question possède, au point a, *les mêmes dérivées que la fonction f jusqu'à l'ordre p inclusivement*, ce qui le caractérise entièrement puisqu'il est de degré p au plus.

La formule de Taylor n'a pas d'intérêt si l'on ne connaît pas de méthode simple pour évaluer l'ordre de grandeur du reste R_p dans (39) ou (40) puisque, dans la pratique, on désire toujours s'en servir pour approcher la fonction f par le polynôme de degré p considéré ci-dessus. Supposons pour cela $a \leqslant b$ dans (38), l'autre cas se traite de même, avec des changements de signe triviaux dans les raisonnements, de sorte que l'on a $(b-x)^p \geqslant 0$ pour $a \leqslant x \leqslant b$. Soit m et M le minimum et le maximum, dans l'intervalle compact $[a, b]$, de la fonction continue $f^{(p+1)}(x)$; on a alors :

$$(42)\quad \begin{aligned} &\int_a^b m\frac{(b-x)^p}{p\,!}dx = m\int_a^b \frac{(b-x)^p}{p\,!}dx \\ &\leqslant R_p \leqslant \int_a^b M\frac{(b-x)^p}{p\,!}dx = M\int_a^b \frac{(b-x)^p}{p\,!}dx, \end{aligned}$$

de sorte que tout revient à évaluer l'intégrale :

$$(43)\quad \int_a^b \frac{(b-x)^p}{p\,!}dx.$$

Pour cela, appliquons la formule (39) à la fonction $f(x) = (x-a)^{p+1}/(p+1)\,!$, dont les dérivées successives sont $(x-a)^p/p\,!$, ..., $x-a$ et 1 pour la dérivée

d'ordre $p+1$; les dérivées d'ordre $\leqslant p$ sont nulles pour $x=a$, de sorte que, pour cette fonction, la formule (39) se réduit à son reste, précisément égal à (43) puisque l'on a, pour la fonction considérée, $f^{(p+1)}(x)=1$ pour tout x ; d'où la valeur de l'intégrale (43), à savoir :

$$(44) \qquad \int_a^b \frac{(b-x)^p}{p!}\,dx = \frac{(b-a)^{p+1}}{(p+1)!}.$$

Portant dans (42) on trouve donc, pour le reste (38) associé à une fonction f de nouveau quelconque, les inégalités :

$$(45) \quad m\,\frac{(b-a)^{p+1}}{(p+1)!} \leqslant \mathrm{R}_p \leqslant \mathrm{M}\,\frac{(b-a)^{p+1}}{(p+1)!},$$

ce qui montre encore que l'on a :

$$(46) \quad \mathrm{R}_p = k\frac{(b-a)^{p+1}}{(p+1)!} \text{ avec } \quad m \leqslant k \leqslant \mathrm{M}.$$

On ne connaît pas exactement le nombre k, mais on sait qu'il est compris entre le minimum et le maximum, sur l'intervalle $[a, b]$, de la fonction continue $f^{(p+1)}(x)$; en appliquant à $f^{(p+1)}$ le théorème 14 qui sera démontré ci-dessous, on en conclut qu'il existe dans l'intervalle $[a, b]$ un point c où l'on a $k=f^{(p+1)}(c)$; portant dans (46) et (39), on obtient finalement le résultat suivant :

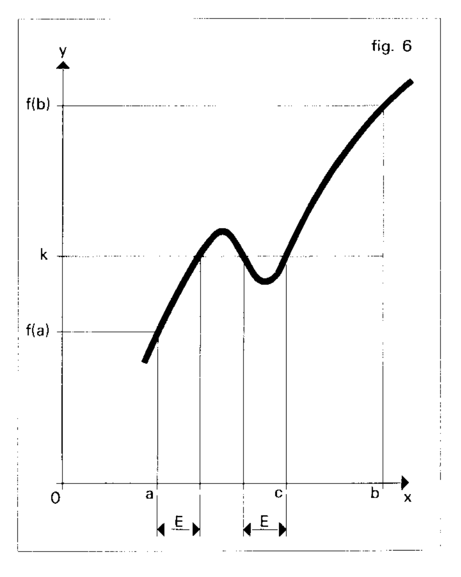
Théorème 13. Soit f une fonction de classe C^{p+1} sur un intervalle X. Quels que soient les points a et b de X, il existe un nombre c compris entre a et b et tel que l'on ait :

$$(47) \quad \begin{aligned} f(b) = f(a) &+ f'(a)(b-a) \\ &+ f''(a)(b-a)^2/2! \\ &+ \ldots + f^{(p)}(a)(b-a)^p/p! \\ &+ f^{(p+1)}(c)(b-a)^{p+1}/(p+1)!. \end{aligned}$$

Il nous reste à établir le théorème 14, auquel nous avons fait allusion plus haut ; c'est, de toute façon, l'une des propriétés les plus importantes des fonctions continues.

Théorème 14. Soit m et M le minimum et le maximum d'une fonction continue sur un intervalle compact $[a, b]$. Pour tout nombre k compris entre m et M, il existe un nombre c compris entre a et b et tel que $k=f(c)$.

On sait (théorème 12) qu'il existe dans $[a, b]$ des points a' et b' où $f(a')=m$ et $f(b')=\mathrm{M}$; remplaçant l'intervalle $[a, b]$ par l'intervalle $[a', b']$, on est encore ramené à l'énoncé suivant (« théorème des valeurs intermédiaires », fig. 6) :



Théorème 14 bis. Soit f une fonction continue sur un intervalle X, a et b deux points de X, et k un nombre compris entre a et b. Il existe un nombre c compris entre a et b tel que $f(c)=k$.

Pour fixer les idées, supposons $a \leqslant b$ et $f(a) \leqslant k \leqslant f(b)$. Soit E l'ensemble des points $x \in [a, b]$ où l'on a $f(x) \leqslant k$; l'ensemble E n'est pas vide (on a visiblement $a \in \mathrm{E}$), et il est borné supérieurement (on a $x \leqslant b$ pour tout $x \in \mathrm{E}$) ; il admet donc (théorème 1) une borne supérieure c. Comme $x \leqslant b$ pour tout $x \in \mathrm{E}$ on a aussi

$c \leqslant b$; comme $a \in E$ on a aussi $a \leqslant c$, de sorte que $c \in [a, b]$. Nous allons montrer que $f(c) = k$, ce qui prouvera le théorème. Il suffit pour cela de montrer que l'on ne peut avoir ni $f(c) > k$, ni $f(c) < k$.

Supposons que l'on ait $f(c) > k$. On aurait alors $a \neq c$ puisque $f(a) \leqslant k$, et donc aussi $a < c$. D'autre part, il existerait un entier p tel que l'on ait encore :

(48) $$f(c) > k + 10^{-p},$$

puisque l'inégalité $f(c) > k$ est stricte. Mais f est continue au point c ; il y a donc un intervalle ouvert $I(c)$ contenant c et tel que f soit constante à 10^{-p} près dans $[a, b] \cap I(c)$. Comme $a < c$, cette intersection contient au moins un c' tel que $a \leqslant c' < c$, donc contient tout l'intervalle $[c', c]$, lequel, puisque $c = \sup(E)$, contient certainement un $x \in E$. On a alors $f(x) \leqslant k$, puisque $x \in E$, et :

$$|f(x) - f(c)| \leqslant 10^{-p},$$

puisque $x \in I(c)$, d'où :

$$f(c) \leqslant f(x) + 10^{-p} \leqslant k + 10^{-p},$$

ce qui contredit (48). L'éventualité dans laquelle on aurait $f(c) > k$ est donc exclue.

Supposons maintenant que l'on ait $f(c) < k$. Comme $k \leqslant f(b)$, on a alors certainement $c < b$. Comme, d'autre part, l'inégalité $f(c) < k$ est stricte, il existe, comme plus haut, un entier p tel que l'on ait :

(49) $$f(c) < k - 10^{-p}.$$

Mais f est continue au point c ; il y a donc, comme plus haut, un intervalle ouvert $I(c)$ contenant c et dans lequel la fonction f est égale, à 10^{-p} près, à $f(c)$. Comme $c < b$, il y a un $c'' \in I(c) \cap [a, b]$ tel que $c < c''$, et, comme on a $|f(c'') - f(c)| \leqslant 10^{-p}$, il vient :

$$f(c'') \leqslant f(c) + 10^{-p} \leqslant k$$

d'après (49). Par suite $c'' \in E$; mais c'est absurde puisque c'' dépasse strictement la borne supérieure c de l'ensemble E. Le théorème 14 *bis* est donc démontré, et, avec lui, la formule de Taylor.

Les applications principales de la formule de Taylor concernent les développements en série des fonctions « élémentaires » : fonctions trigonométriques, exponentielles, logarithmiques, etc. (cf. EXPONENTIELLE ET LOGARITHME, FONCTIONS ANALYTIQUES). On trouvera également, ci-après, l'extension de la formule de Taylor aux fonctions de plusieurs variables.

ROGER GODEMENT

Bibliographie

H. BEHNKE, F. BACHMANN et al., *Fundamentals of Mathematics*, vol. III : *Analysis*, M.I.T. Press, Cambridge, 1984 / N. BOURBAKI, *Fonctions d'une variable réelle*, Masson, Paris, 1982 / R. COUTY & J. EZRA, *Analyse*, Armand Colin, 5e éd. 1980 / A. DELEDICQ & M. DIENER, *Leçons de calcul infinitésimal*, *ibid.*, 1989 / J. DIEUDONNÉ, *Éléments d'analyse*, t. I : *Fondements de l'analyse moderne*, Gauthier-Villars, 3e éd. 1979 ; *Calcul infinitésimal*, Hermann, 2e éd. 1980 / J. DIXMIER, *Cours de mathématiques du premier cycle*, 2 vol., Gauthier-Villars, Paris, 2e éd., 1976-1977 / C. DOUCHET & B. ZWAHLEN, *Calcul différentiel et intégral*, vol. I et III : *Fonctions réelles d'une variable réelle*, Presses polytechniques romandes, Lausanne, 1983 / H. GRAUERT & I. LIEB, *Differential- und Integralrechnung*, vol. I, Springer-Verlag, Berlin, 1976 / S. LANG, *Analyse réelle*, InterÉditions, 1977 / P. LAX, S. BURSTEIN & A. LAX, *Calculus with Applications and Computing*, Springer, New York, 1983 / J. PICHON, *Calcul des limites*, Ellipses, 1987 / O. TOEPLITZ, *Die Entwicklung der Infinitesimalrechnung*, Berlin, 1950 ; trad. angl., *The Calculus, a Genetic Approach*, Univ. of Chicago Press, Chicago, 1981.

B. Calcul à plusieurs variables

Le calcul infinitésimal des fonctions de plusieurs variables a eu un développement

plus tardif que celui des fonctions d'un seul argument. Inauguré avec un siècle de retard, il ne parvient à établir solidement ses fondements qu'au début du XXᵉ siècle.

Ce n'est qu'aux environs de 1930 que sont abordés les problèmes difficiles de cette branche de l'analyse, très utilisée depuis lors.

1. La préhistoire

Le formalisme des dérivées partielles

Avant d'étudier le comportement d'une fonction $f(x,y)$ de deux variables, lorsque x et y varient *simultanément* et *indépendamment*, on commence par faire varier x et y *successivement*. Fixons la valeur de y : la dérivée de la fonction $x \mapsto f(x,y)$, lorsqu'elle existe, s'appelle la dérivée partielle $\frac{\partial f}{\partial x}(x,y)$ de f par rapport à x (à y constant). La notation utilisant le ∂ pour désigner la dérivation partielle, par opposition au d désignant la dérivation ordinaire, a été préconisée par Legendre (1786) et vulgarisée par Jacobi (1841). Si, maintenant, on fait varier x et y en fonction d'une même variable t, on trouve que :

$$(1) \quad \frac{df}{dt}[x(t), y(t)] = \frac{dx}{dt}(t)\cdot\frac{\partial f}{\partial x}[x(t), y(t)] + \frac{dy}{dt}(t)\cdot\frac{\partial f}{\partial y}[x(t), y(t)],$$

ce qui fait apparaître les dérivées partielles de f comme des intermédiaires de calcul commodes.

Les dérivées partielles apparaissent, en 1755, dans le traité *Institutiones calculi differentialis* d'Euler, et, en 1747, chez A. Clairaut. Ils y ont reconnu l'outil de base du calcul différentiel à plusieurs variables. Malheureusement, cette notion est essentiellement liée au choix d'un système de coordonnées.

Par exemple, considérons les formules $W = RI^2 = EI = E^2/R$, qui traduisent un cas particulier des lois d'Ohm en électricité. On constate que le symbole $\partial W/\partial I$ est égal à 2 RI, E ou 0 selon l'expression de W que l'on adopte. L'explication de ce paradoxe vient de ce que la dérivation par rapport à I n'a pas la même signification selon que l'on opère à R constant (et E variable), ou à E constant (et R variable) et enfin si l'on fixe E et R.

S'inspirant de ce que Leibniz avait fait pour la différentielle des fonctions d'une variable, Euler et Clairaut étudièrent une expression remarquable, la *différentielle totale* :

$$(2) \qquad df = \frac{\partial f}{\partial x}(x,y)dx + \frac{\partial f}{\partial y}(x,y)dy\,;$$

c'est une fonction de quatre variables indépendantes x, y, dx, dy (linéaire par rapport aux deux dernières). Cette expression possède un caractère invariant lorsqu'on la soumet à des changements de variables du type particulier suivant : si l'on exprime x et y en fonction de nouvelles variables X,Y au moyen des formules $x = x(X,Y)$, $y = y(X,Y)$, et si l'on effectue, *en même temps*, le changement de variables « covariant », selon les formules :

$$dx = \frac{\partial x}{\partial X}(X,Y)dX + \frac{\partial x}{\partial Y}(X,Y)dY,$$
$$dy = \frac{\partial y}{\partial X}(X,Y)dX + \frac{\partial y}{\partial Y}(X,Y)dY,$$

on constate que la différentielle totale de f prend la même forme (2) que l'on exprime df à l'aide de x, y, dx et dy ou à l'aide de X, Y, dX et dY.

Les dérivées partielles sont susceptibles d'être dérivées partiellement à leur tour.

On obtient ainsi les dérivées partielles secondes :

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right), \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right), \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right), \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right),$$

que Jacobi note respectivement :

$$\frac{\partial^2 f}{\partial x^2}, \quad \frac{\partial^2 f}{\partial y^2}, \quad \frac{\partial^2 f}{\partial y\,\partial x}, \quad \frac{\partial^2 f}{\partial x\,\partial y}.$$

Euler et Clairaut avaient déjà constaté que le résultat ne dépend pas de l'ordre dans lequel on effectue les dérivations (pour autant que ces auteurs n'opéraient tacitement que sur de « bonnes fonctions », c'est-à-dire des fonctions que nous nommons aujourd'hui analytiques), soit :

$$(3) \qquad \frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 f}{\partial y\,\partial x}.$$

À titre de preuve, Euler se borne à invoquer la symétrie de l'expression :

$$f(x+h, y+k) - f(x,y+k) - f(x+h,y) + f(x,y),$$

et Clairaut opère formellement sur un développement en série.

Une combinaison linéaire à coefficients complexes de dérivations partielles s'appelle un opérateur différentiel linéaire à coefficients constants (par exemple :

$$\mathrm{A}\frac{\partial}{\partial x} + \mathrm{B}\frac{\partial^2}{\partial y^2} - \mathrm{C}\frac{\partial^3}{\partial x\,\partial y^2},$$

où A, B, C sont des nombres complexes). Ces opérateurs sont susceptibles d'être ajoutés et composés entre eux : il en résulte une structure algébrique sur l'ensemble des opérateurs différentiels à coefficients constants qu'en langage moderne on peut formuler ainsi : « L'ensemble des opérateurs différentiels à coefficients complexes opérant sur des fonctions de deux variables est un anneau (pour l'addition et la composition), isomorphe à l'anneau des polynômes C [X,Y] à deux indéterminées X et Y. » (Par exemple, à l'opérateur cité ci-dessus correspond le polynôme $AX + BY^2 - CXY^2$.) L'isomorphisme précédent s'explique par l'analogie complète qui existe entre les règles de calcul qui régissent l'addition et la multiplication des polynômes d'une part, l'addition et la composition des opérateurs différentiels à coefficients constants d'autre part. Parmi ces règles, la formule (3) exprime la commutativité des dérivations partielles. Mais, au XVIIIe siècle, de telles explications ne pouvaient être pleinement comprises : les calculateurs imaginaient une profusion de règles de calcul « symboliques », dont l'efficacité restait mystérieuse.

Voici une application des considérations précédentes. En tenant compte de l'isomorphisme précité, on peut développer l'opérateur $(\partial/\partial x + \partial/\partial y)$ itéré n fois, grâce à la formule du binôme de Newton :

$$\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^{(n)} = \sum_{k=0}^{k=n} C_n^k \frac{\partial^k}{\partial x^k} \cdot \frac{\partial^{n-k}}{\partial y^{n-k}} \cdot$$

En appliquant les deux membres à la fonction $f(x)g(y)$ et en remplaçant y par x dans le résultat, on aboutit à la formule de Leibniz :

$$\frac{d^n}{dx^n}[f(x)g(x)] = \sum_{k=0}^{k=n} C_n^k \frac{d^k f}{dx^k}(x) \frac{d^{n-k} g}{dx^{n-k}}(x).$$

La théorie des équations aux dérivées partielles oblige à manipuler de plus en plus des expressions différentielles : l'importance du laplacien : $\partial^2/\partial x^2 + \partial^2/\partial y^2$ provient du fait que c'est (à un coefficient près) le seul opérateur différentiel à coefficients constants, homogène du second ordre, qui soit invariant par rotation des axes. Le laplacien, à cause de cette signification

intrinsèque, apparaît dans de nombreux phénomènes physiques.

L'étude de certaines équations aux dérivées partielles liées à des problèmes de géométrie différentielle a conduit à adopter les notations de Monge :

$$p = \frac{\partial f}{\partial x}, \quad q = \frac{\partial f}{\partial y}, \quad r = \frac{\partial^2 f}{\partial x^2}, \quad s = \frac{\partial^2 f}{\partial x \partial y},$$
$$t = \frac{\partial^2 f}{\partial y^2},$$

utilisées à la fin du XIXe siècle, mais qui sont tombées en désuétude, faute de pouvoir se généraliser aux cas des fonctions de n variables, et à valeurs vectorielles.

Euler s'est également occupé du calcul des dérivées d'une *fonction implicite* : partant d'une équation $f(x,y) = 0$, il suppose, sans trop de scrupules, que l'on peut trouver une fonction $x \mapsto y(x)$ satisfaisant à $f[x,y(x)] \equiv 0$. Il calcule alors la dérivée $y'(x)$ à partir de l'équation :

$$\frac{\partial f}{\partial x}[x,y(x)] + \frac{dy}{dx}\frac{\partial f}{\partial y}[x,y(x)] = 0,$$

et en déduit l'expression des dérivées successives de y.

Étudiant de la même façon la résolution d'un système :

$$\begin{cases} f_1(x,y_1,y_2,y_3) = 0 \\ f_2(x,y_1,y_2,y_3) = 0 \\ f_3(x,y_1,y_2,y_3) = 0 \end{cases}$$

Jacobi introduit la notion de déterminant fonctionnel (appelé aussi déterminant jacobien) défini par :

$$\det \begin{vmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_2}{\partial y_1} & \frac{\partial f_3}{\partial y_1} \\ \frac{\partial f_1}{\partial y_2} & \frac{\partial f_2}{\partial y_2} & \frac{\partial f_3}{\partial y_2} \\ \frac{\partial f_1}{\partial y_3} & \frac{\partial f_2}{\partial y_3} & \frac{\partial f_3}{\partial y_3} \end{vmatrix}$$

et dégage les règles du calcul formel qui le régissent. Il met en évidence la notion de dépendance fonctionnelle analogue à la dépendance linéaire entre formes linéaires.

Au XIXe siècle commence l'étude algébrique des opérateurs différentiels linéaires à coefficients variables : deux tels opérateurs ne commutent pas nécessairement (on s'en convaincra en composant les opérateurs d/dx et $x\,(d/dx)$ relatifs à des fonctions d'une seule variable). On introduit alors diverses expressions différentielles (déterminants hessiens ou wronskiens, crochets de Lie, parenthèses de Poisson, invariants intégraux, etc.), qui préparent la voie au calcul différentiel extérieur, à l'analyse tensorielle et à la géométrie différentielle moderne (cf. GÉOMÉTRIE DIFFÉRENTIELLE CLASSIQUE). L'aboutissement de cet effort d'algébrisation sera l'édification d'une science traitant abstraitement des objets mathématiques soumis aux mêmes règles de calcul que les opérateurs différentiels : l'*algèbre différentielle* ainsi édifiée s'affranchit de toutes considérations de continuité et de limite, et trouve des applications dans l'étude des fonctions définies sur un corps quelconque et jusqu'en arithmétique.

En ce qui concerne les notations, un perfectionnement important a été apporté, en 1934, par H. Whitney, qui a introduit l'usage des *multi-indices*. Un multi-indice à n variables est un système ordonné de n entiers $(\alpha_1, \alpha_2, ..., \alpha_n)$, que l'on désigne par un seul symbole α. Dans ces conditions, on écrit :

$$\frac{\partial^\alpha f}{\partial X^\alpha} \quad \text{au lieu de} \quad \frac{\partial^{\alpha_1 + \alpha_2 + ... + \alpha_n}}{\partial X_1^{\alpha_1} \partial X_2^{\alpha_2} \partial X_n^{\alpha_n}},$$
$$X^\alpha \text{ au lieu de } X_1^{\alpha_1} X_2^{\alpha_2} ... X_n^{\alpha_n} ;$$

on pose, de plus :

$$|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_n \quad \text{et} \quad \alpha! = \alpha_1!\,\alpha_2! ... \alpha_n! ;$$

avec de telles notations, la série de Taylor s'écrit alors comme une série à une variable, mais avec une sommation étendue à l'ensemble de tous les multi-indices :

$$\sum \frac{(X-X_0)^\alpha}{\alpha\,!} \cdot \frac{\partial^\alpha f}{\partial X^\alpha}(X_0).$$

Formulation intrinsèque de la théorie

Les inconvénients des dérivées partielles posèrent, dès l'apparition du calcul vectoriel, le problème de la formulation intrinsèque de la théorie, en mettant en évidence des expressions invariantes par changement de coordonnées ; M étant un point de coordonnées $(x,y,z, \ldots)$, on ne parlera plus de fonctions des variables, mais de fonctions du point M.

Une étape historique importante, aujourd'hui complètement dépassée, a été l'élaboration, par O. Heaviside et W. Gibbs, de l'analyse vectorielle, qui met l'accent sur certaines expressions invariantes par changement de coordonnées rectangulaires dans l'espace à trois dimensions. La structure euclidienne y joue donc un rôle primordial. À une fonction numérique $M \mapsto f(M) = f(x,y,z)$ (où x, y, z sont les coordonnées de M par rapport à une base orthonormée), on associe le vecteur, dont les composantes sont :

$$\frac{\partial f}{\partial x}(M), \frac{\partial f}{\partial y}(M), \frac{\partial f}{\partial z}(M).$$

C'est le gradient de la fonction f. Il est lié à la différentielle totale de f par la formule :

$$df = \overrightarrow{\mathrm{grad}} f \,.\, d\,\vec{M},$$

où le second membre représente un produit scalaire et $d\vec{M}$ le vecteur dont les composantes sont dx, dy, dz. À un champ de vecteur $M \mapsto \vec{V}(M)$, défini par les trois composantes X(M), Y(M), Z(M) du vecteur $\vec{V}$, on associe une fonction scalaire, la *divergence* du champ :

$$\mathrm{div}\,\vec{V} = \frac{\partial X}{\partial x} + \frac{\partial Y}{\partial y} + \frac{\partial Z}{\partial z}$$

et une fonction vectorielle, le *rotationnel* de $\vec{V}$, dont les composantes sont :

$$\overrightarrow{\mathrm{rot}}\,(\vec{V}) = \left(\frac{\partial Z}{\partial y} - \frac{\partial Y}{\partial z}, \frac{\partial X}{\partial z} - \frac{\partial Z}{\partial x}, \frac{\partial Y}{\partial x} - \frac{\partial X}{\partial y}\right).$$

Le gradient, la divergence et le rotationnel sont des notions invariantes par changement d'axes orthonormés et sont soumises à un grand nombre de règles de calcul « symboliques », dont les plus simples sont :

$$\overrightarrow{\mathrm{rot}}\,(\overrightarrow{\mathrm{grad}} f) = 0, \quad \mathrm{div}\,(\overrightarrow{\mathrm{rot}}\,\vec{V}) = 0.$$

Ces notions permettent d'écrire, sous une forme concise et suggestive, un grand nombre de formules de la physique théorique, de la mécanique et de la géométrie. Par exemple, les divers cas particuliers de la formule de Stokes (attribués à Green, Ampère, Ostrogradsky, etc.) prennent la forme :

$$\iiint \overrightarrow{\mathrm{grad}} f \,.\, d\tau = \iint f \,.\, \vec{n}\, d\sigma,$$

$$\iint \mathrm{div}\,\vec{V} \,.\, d\tau = \iint (\vec{V} \,.\, \vec{n})\, d\sigma,$$

$$\iiint \overrightarrow{\mathrm{rot}}\,\vec{V} \,.\, d\tau = \iint \vec{n} \wedge \vec{V}\, d\sigma,$$

où $d\tau$ (resp. $d\sigma$) désigne l'élément de volume (resp. d'aire), où l'intégrale triple est étendue à un domaine tridimensionnel orienté, et où l'intégrale de surface est étendue au bord orienté dans ce domaine, $\vec{n}$ représente le vecteur normal à la surface considérée.

Mais cette analyse vectorielle ne couvre pas l'ensemble de tous les besoins de la physique mathématique. Elle ne déborde pas le cadre des fonctions de trois variables

et des changements de bases orthonormales. La véritable solution du problème de la formulation intrinsèque du calcul différentiel n'a été obtenue qu'après l'édification de l'analyse tensorielle (par Ricci et Levi-Civita) et du calcul différentiel extérieur (par Élie Cartan). Bien que ces théories n'aient été exposées à l'origine qu'en termes de coordonnées, il n'a pas été difficile (après la construction axiomatique de l'algèbre linéaire et multilinéaire), de les traduire en langage intrinsèque.

Le calcul différentiel extérieur explique l'origine (qui paraissait quelque peu mystérieuse) des notions de gradient, de rotationnel et de divergence, puisque les dérivées extérieures de :

$$f(x, y, z),$$
$$X(x, y, z)\,dx + Y(x, y, z)\,dy + Z(x, y, z)\,dz,$$
$$X(x, y, z)\,dy \wedge dz + Y(x, y, z)\,dz \wedge dx + Z(x, y, z)\,dx \wedge dy$$

sont respectivement :

$$\frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz,$$
$$\left(\frac{\partial Z}{\partial y} - \frac{\partial Y}{\partial z}\right)dy \wedge dz + \left(\frac{\partial X}{\partial z} - \frac{\partial Z}{\partial x}\right)dz \wedge dx + \left(\frac{\partial Y}{\partial x} - \frac{\partial X}{\partial y}\right)dx \wedge dy,$$
$$\left(\frac{\partial X}{\partial x} + \frac{\partial Y}{\partial y} + \frac{\partial Z}{\partial z}\right)dx \wedge dy \wedge dz.$$

La formule de Stokes qui s'exprimait, nous l'avons vu, sous des formes très diverses et d'emploi limité dans le langage de l'analyse vectorielle s'écrit aujourd'hui :

$$\int_{\partial\Omega} \omega = \int_{\Omega} d\omega,$$

où ω est une forme différentielle, $d\omega$ sa différentielle extérieure, Ω une chaîne orientée, et $\partial\Omega$ le bord orienté de Ω.

Les progrès de l'algèbre linéaire ont permis enfin de définir la différentielle sans aucun recours aux coordonnées sous une forme qui s'applique également aux fonctions définies sur des espaces de dimension infinie (cf. chap. 2).

Après que R. Gateaux et V. Volterra eurent dégagé la notion de dérivée directionnelle d'une fonction définie sur un espace fonctionnel (établissant ainsi la synthèse entre le calcul des variations [cf. calcul des VARIATIONS] et le calcul différentiel classique), O. Stoltz et M. Fréchet donnaient la définition intrinsèque de la notion de différentielle. Ces travaux ont fait prendre conscience du fait fondamental suivant : pour traiter le calcul différentiel des fonctions définies sur un espace vectoriel normé E et prenant leurs valeurs dans un espace normé F, il est indispensable de manier simultanément de nombreux espaces auxiliaires obtenus à partir de produits tensoriels des espaces E, F et du dual de E. En particulier, chacune des dérivées successives d'une application différentiable *f* doit prendre ses valeurs dans un espace différent.

Cette conception oblige à manipuler simultanément un grand nombre d'espaces normés, ayant des dimensions différentes, ce qui n'est pas fait pour faciliter l'intuition géométrique. Pour obvier à cet inconvénient, C. Ehresmann a créé, en 1943 et en 1952, deux outils extrêmement commodes : la notion d'*espace fibré* et la notion de *variété des jets.* Grâce à ce langage, on peut concevoir le support géométrique du calcul différentiel à n variables comme un domaine Ω de l'espace à n dimensions, et, « au-dessus » de chaque point M de Ω, on imagine une « fibre », formée par une collection d'objets mathématiques (vecteurs, tenseurs, matrices, covecteurs et cotenseurs, etc.), dont on aura besoin pour décrire la théorie. La description a priori de la variété des jets, exprimée de façon intrinsèque, permet de s'affranchir de la manipulation de signes hérissés d'indices

compliqués et d'utiliser à nouveau un langage géométrique.

Apparition de la rigueur

Le flux de formalisme qui caractérise la première époque du calcul infinitésimal fut suivi par un reflux critique dans le domaine de la rigueur.

Sous l'influence de Cauchy, Abel et Weierstrass, on se préoccupa de reprendre les principales formules découvertes plus ou moins empiriquement au siècle précédent, pour en préciser au mieux les limites de validité. Ainsi, H. A. Schwarz prouva, en 1873, que la formule :

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

est valable, sous réserve de la continuité d'un des deux membres par rapport à l'ensemble des variables. Peano donna l'exemple de la fonction :

$$f(x,y) = xy\frac{x^2 - y^2}{x^2 + y^2},$$

prolongée par continuité en posant $f(0,0) = 0$, pour laquelle la permutation des dérivées partielles n'est pas licite.

Peano entreprit systématiquement de dépister les affirmations non rigoureuses, largement répandues à l'époque, construisit des contre-exemples, aujourd'hui classiques, et tenta de redémontrer certains théorèmes, sous les hypothèses les plus faibles possibles. C'est ainsi qu'il améliora l'exposition du théorème des accroissements finis et démontra la formule de Taylor (cf. chap. 2), pour les fonctions d'une ou plusieurs variables, en éliminant toute hypothèse superflue ; on lui doit en outre la formule classique sur le reste de Young (trouvée par Peano avant cet auteur). Il signala une erreur célèbre touchant les maxima (ou minima) des fonctions : on pensait que l'inspection des termes de plus bas degré du développement de Taylor, au voisinage de l'origine d'une fonction, ainsi que le comportement de cette fonction sur chaque droite passant par l'origine permettait de décider de l'existence d'un extrémum en ce point. Or la fonction $(y - x^2)(y - 2x^2)$, pour laquelle le terme de plus bas degré est y^2, admet un minimum relatif nul sur chaque droite passant par l'origine ; cependant, elle prend des valeurs négatives dans tout voisinage de l'origine.

Vers la même époque paraissent, dans les traités d'analyse, les premières démonstrations correctes du théorème des fonctions implicites et de quelques-unes de ses variantes. La préhistoire prend fin, le véritable développement commencera trente ans plus tard.

2. Exposé moderne de la théorie élémentaire

Dérivée première

Soit E et F deux espaces normés, et Ω un ensemble ouvert de E : on dit que deux fonctions continues f et g (définies sur Ω et à valeurs dans F) *admettent un contact d'ordre r* (où r est un nombre entier) *au point* $A \in \Omega$ si le rapport :

$$\frac{\|f(M) - g(M)\|}{\|AM\|^r}$$

tend vers 0 lorsque M tend vers A. En particulier, lorsque $r = 1$ on dit que f et g sont *tangentes au point* A ; cette définition implique que $f(A) = g(A)$.

Une fonction continue f (définie sur Ω et à valeurs dans F) *est dérivable* en $A \in \Omega$, s'il existe une fonction continue affine :

$$M \mapsto f(A) + L[\overrightarrow{AM}]$$

(où L est une application linéaire continue de E dans F, c'est-à-dire un élément de $\mathcal{L}$(E,F) qui est tangente à f au point A). L s'appelle aujourd'hui la *dérivée* de f au point A (dans l'ancienne terminologie, c'était la « *différentielle au sens de Stolz-Fréchet* » ou plus brièvement la *différentielle* de f au point A). On la note d'ordinaire $D^1f(A)$. Lorsque f est dérivable en tout point de Ω, la fonction dérivée est l'application $A \mapsto D^1f(A)$ définie dans Ω et à valeurs dans $\mathcal{L}$(E,F). On dit que f est *continûment dérivable* (ou encore de *classe* C^1) si la fonction dérivée est continue, lorsqu'on munit $\mathcal{L}$(E,F) de sa norme usuelle :

$$\|L\| = \sup_{\|x\| \leqslant 1} \|L(x)\|.$$

Dans le cas particulier où $E = \mathbf{R}^n$ et $F = \mathbf{R}$, toute fonction de classe C^1 admet des dérivées partielles continues et la dérivée de f au point de coordonnées $(x_1, x_2, ..., x_n)$ est la forme linéaire qui associe au vecteur de coordonnées $(dx_1, dx_2, ..., dx_n)$ le nombre :

$$\frac{\partial f}{\partial x_1}(x_1, x_2, ..., x_n)\, dx_1 + \frac{\partial f}{\partial x_2}(x_1, ..., x_n)\, dx_2 + ... + \frac{\partial f}{\partial x_n}(x_1, ..., x_n)\, dx_n\ ;$$

la dérivée coïncide donc avec la « différentielle totale » d'Euler (cf. chap. 1).

Lorsque $E = \mathbf{R}^n$ et $F = \mathbf{R}^p$ la fonction f est définie par p fonctions numériques f_1, f_2,..., f_p. La dérivée de f en un point de coordonnées $(x_1, x_2, ..., x_n)$ est, lorsqu'elle existe, l'application linéaire appartenant à $\mathcal{L}(\mathbf{R}^n, \mathbf{R}^p)$ définie par la matrice jacobienne des fonctions f_j ($j \leqslant p$) par rapport aux variables x_i ($i \leqslant n$).

Une fonction qui possède des dérivées partielles en chaque point de Ω n'est pas nécessairement dérivable, comme le montre l'exemple de la fonction scalaire :

$$f(x,y) = \frac{xy}{\sqrt{x^2 + y^2}}$$

(prolongée par $f(0, 0) = 0$). Son graphe est un demi-cône dont le sommet est à l'origine ; il ne possède donc pas de plan tangent en ce point. Par contre, on montre que toute fonction qui admet des dérivées partielles continues par rapport à l'ensemble des variables est nécessairement dérivable.

Dérivées successives

Supposons que la fonction dérivée $M \mapsto D^1f(M)$ soit elle-même dérivable : sa dérivée $D^1(D^1f)(M)$ appartient à $\mathcal{L}[E, \mathcal{L}(E,F)]$. On peut l'identifier à une application bilinéaire continue de $E \times E$ dans F, et dans ce cas on l'appelle la dérivée seconde de f au point M et on la note $D^2f(M)$. On peut ainsi définir les dérivées successives de proche en proche.

On peut également donner une définition directe des fonctions r fois continûment dérivables (ou de classe $\mathbf{C}^r$) en utilisant la notion de *polynôme* défini sur E et à valeurs dans F. Désignons par $\mathcal{L}_k$ (E,F) l'espace vectoriel des applications continues k-linéaires, symétriques, définies sur $(E)^k$ et à valeurs dans F. En d'autres termes $L_k \in \mathcal{L}_k(E, F)$ est une application :

$$(\vec{V}_1, \vec{V}_2, ..., \vec{V}_k) \rightarrow L_k[\vec{V}_1, \vec{V}_2, ..., \vec{V}_k]$$

séparément linéaire par rapport à chaque argument, invariante lorsqu'on permute arbitrairement les vecteurs V_i entre eux, et telle que $\|L_k[\vec{V}_1, \vec{V}_2, ..., \vec{V}_k]\|_F$ reste borné lorsque les vecteurs $\vec{V}_i$ restent tous dans la boule unité de E. Un *monôme de degré k* défini dans E et à valeurs dans F est alors par définition une fonction de point :

$$M \mapsto L_k[\overrightarrow{OM}, \overrightarrow{OM}, ..., \overrightarrow{OM}]\ ;$$

remarquons que lorsque E est de dimension finie et que l'on exprime $\overrightarrow{OM}$ à l'aide de ses composantes, on n'obtient pas un monôme ordinaire, mais un polynôme *homogène* de degré k ayant pour coefficients des vecteurs de F. Un *polynôme* sur E et à valeurs dans F est maintenant, par définition, une somme finie de monômes.

Une formule de Taylor permet de remplacer les vecteurs $\overrightarrow{OM}$ qui interviennent dans la définition précédente par des vecteurs $\overrightarrow{AM}$ où $A \in E$. On dira qu'une fonction f, définie sur Ω, admet un *développement limité d'ordre m* au point A, s'il existe des $L_k \in \mathcal{L}_k(E, F)$, $k = 1,2, ..., n$, tels que le polynôme :

$$T_A f(B) = f(A) + \sum_{k=1}^{m} L_k [\overrightarrow{AB}, \overrightarrow{AB}, ..., \overrightarrow{AB}]$$

satisfasse à la condition :

$$(1)\ \|f(B) - T_A f(B)\|_F \leqslant \frac{\|AB\|^m}{m\,!} o_A(\|\overrightarrow{AB}\|),$$

où la fonction o_A tend vers 0 lorsque $\|\overrightarrow{AB}\|$ tend vers 0.

Si un tel « polynôme de Taylor » existe en tout point $A \in \Omega$, si les « coefficients » $k\,!L_k$ (que l'on note plutôt $D^k f(A)$ et que l'on appelle les *dérivées k-ièmes de f*) varient continûment sur Ω, et s'il existe en chaque point $A \in \Omega$ une fonction o_A satisfaisant à la condition (1) on dit que f est de classe C^m dans Ω (ou encore m fois continûment dérivable). Il est possible alors de transposer au cas des fonctions définies sur Ω la plupart des formules à une variable. Par exemple, si f est de classe C^{m+1}, le reste de son développement d'ordre m, qui s'écrit, dans le cas d'une variable :

$$\int_a^b \frac{(b-x)^m}{m\,!} f^{(m+1)}(x)dx,$$

s'exprime ici par :

$$(2)\ \int_{AB} \frac{1}{m\,!} D^{m+1} f(X) (\underbrace{\overrightarrow{XB}, \overrightarrow{XB}, ..., \overrightarrow{XB}}_{m \text{ fois}}, \overrightarrow{dX})$$

où l'intégrale curviligne est prise le long de n'importe quelle courbe rectifiable joignant A à B dans Ω.

L'expression (2), appelée *reste intégral* de la formule de Taylor, donne la valeur exacte de $f(B) - T_A f(B)$. On utilise aussi diverses majorations de ce reste. Si la dérivée $A \rightarrow D_A^m f$ est une fonction uniformément continue dans Ω, on peut majorer le reste par une expression de la forme :

$$\frac{\|AB\|^m}{m\,!} \omega(\|\overrightarrow{AB}\|),$$

où $\omega(x)$ est une fonction continue qui tend vers 0, lorsque x tend vers 0 (*majoration de Young*).

Le théorème des fonctions implicites et ses variantes

Étant donné deux domaines $\Omega \subset E$ (resp. $\Omega_1 \subset E_1$), rappelons qu'une application f de Ω sur Ω_1 est un *homéomorphisme* si f est bijective, continue ainsi que l'application réciproque f^{-1}. Un homéomorphisme peut être de classe C^m mais on dit que c'est un *difféomorphisme* (de classe C^m) si l'application réciproque f^{-1} est également de classe C^m (et l'on montre qu'il suffit pour cela que f^{-1} soit de classe C^1). L'exemple de l'application $t \mapsto t^3$ de **R** sur **R** montre qu'un homéomorphisme de classe C^1 n'est pas nécessairement un difféomorphisme.

Théorème d'inversion locale. E et F étant deux espaces de Banach (c'est-à-dire des espaces normés *complets*), soit f une application de classe C^m définie dans un voisinage du point $A \in E$ et prenant ses valeurs dans F. Si la dérivée $D^1 f(A)$ est un isomorphisme linéaire de E sur F, alors il

existe un voisinage U de A dans E et un voisinage U_1 de $f(A)$ dans F, tel que la restriction de f à U soit un difféomorphisme sur U_1.

Commentons cet énoncé. Une fonction différentiable f est « approximativement » égale au voisinage de A à sa dérivée $D^1f(A)$. Le théorème précédent exprime que si l'application linéaire tangente en A est inversible alors f est lui-même inversible au voisinage de A. Ainsi le comportement de la dérivée en un point conditionne le comportement local de f dans tout un voisinage de A.

Le théorème précédent s'applique, par exemple, à l'application de $\mathbf{R}^2$ dans $\mathbf{R}^2$ définie par $X = e^x \cos y$, $Y = e^x \sin y$ (inspirée par l'application $z \mapsto e^z$ de C dans C). On notera que cette application n'est pas globalement bijective.

Théorèmes de submersion et d'immersion locale. Si l'un au moins des espaces de Banach E et F est de dimension finie, soit f une application de classe C^m, définie dans un voisinage de $A \in E$. Si la dérivée $D^1f(A)$ est *surjective* (resp. *injective*), alors il existe un voisinage U de A dans E et un voisinage U_1 de $f(A)$ dans F, et enfin un difféomorphisme θ de U sur U (resp. un difféomorphisme θ_1 de U_1 sur U_1) tel que $f \circ \theta$ (resp. $\theta_1 \circ f$) soit la restriction à U d'une application linéaire surjective (resp. injective) de E dans F.

Ces théorèmes signifient que, si une application f est « approximativement » une application linéaire surjective (resp. injective), on peut la transmuer en une application linéaire du même type. La restriction sur la dimension finie d'un des espaces n'est pas indispensable, mais il faut ajouter une hypothèse supplémentaire pour que les énoncés précédents restent valables.

Forme archaïque du théorème de submersion locale (théorème des fonctions implicites). Si p fonctions :

$$f_j(x_1, x_2, ..., x_{n-p}, y_1, y_2, ..., y_p)$$

(où $j = 1, 2, ..., p$) sont de classe C^m et si le déterminant jacobien des fonctions f_j par rapport aux variables y_j ne s'annule pas à l'origine, on peut exprimer localement les y_j comme fonctions de classe C^m des variables $X = (x_1, x_2, ..., x_{n-p})$ de façon à satisfaire au système d'équations « implicites » :

$$f_j[X, y_1(X), y_2(X), ..., y_p(X)] = 0$$

(où $j = 1, 2, ..., p$).

Il existe une autre variante de ces théorèmes, appelée *théorème du rang constant*, relative au cas où la dérivée de l'application f garde un rang constant au voisinage de l'origine.

Les « images » et « noyaux » des applications différentiables satisfaisant aux énoncés précédents sont localement des morceaux de « variétés différentiables », qui devront être convenablement recollés pour aboutir à une théorie globale.

Un théorème beaucoup plus fin, démontré par J. Nash (1956), concerne le cas où $D^1_A f$, sans être surjective, a une image partout dense dans l'espace de Banach F, de dimension infinie. Dans ces conditions, la résolution locale du système d'équations implicites ne peut s'obtenir qu'au prix d'une certaine perte de dérivabilité.

3. La théorie fine contemporaine

L'œuvre de H. Whitney

Le mathématicien américain Hassler Whitney, dont la contribution à des branches très variées des mathématiques a été sou-

vent décisive (théorie des graphes, topologie algébrique et différentielle, axiomatisation de la notion de variété ou de produit tensoriel, étude des ensembles analytiques réels, etc.), est le véritable initiateur du renouveau du calcul différentiel des fonctions de plusieurs variables. Parmi ses contributions, citons :

Le théorème du prolongement (1934). En chaque point d'un ensemble fermé F quelconque de $\mathbf{R}^n$ on se donne un polynôme à n variables de degré inférieur ou égal à m. Whitney énonce une condition nécessaire et suffisante pour qu'un tel *champ de polynômes* soit la restriction à F du champ des polynômes de Taylor d'une fonction f de classe C^m. Ce théorème permet, en particulier, de construire des fonctions f de classe C^m ayant certaines singularités données à l'avance : on commence par se donner l'ensemble F des points singuliers puis l'on cherche à prolonger un champ de polynômes défini sur F. Indiquons que ce théorème est encore valable pour $m = \infty$. Un cas particulier très important est celui où F se réduit à un seul point ; on obtient le *théorème d'É. Borel généralisé à n variables* : « Il existe une fonction f de classe C^∞ dont les dérivées partielles prennent en un point donné des valeurs arbitrairement choisies. » En d'autres termes, la série de Taylor d'une fonction de classe C^∞ peut être n'importe quelle *série formelle*.

Caractérisation des idéaux fermés de fonctions différentiables. L'ensemble des fonctions numériques de classe C^m définies sur un pavé compact K constitue une algèbre de Banach (lorsque m est fini) pour la norme de la *convergence uniforme d'ordre m* (c'est-à-dire de la convergence uniforme de chacune des dérivées partielles). La structure des idéaux de cette algèbre est d'une grande complexité, mais H. Whitney a démontré, en 1944, le théorème suivant concernant les idéaux *fermés*.

Théorème. Pourqu'une fonction f de classe C^m sur K appartienne à un idéal *fermé* $\mathfrak{I}$, il faut et il suffit qu'à chaque point de K on puisse associer une fonction $g_A \in \mathfrak{I}$ telle que f et g_A aient les mêmes polynômes de Taylor en A. Ce théorème signifie que l'appartenance *ponctuelle à* $\mathfrak{I}$ implique l'appartenance *globale*.

Classification des singularités

En 1925, le mathématicien américain Marston Morse a inauguré l'étude des singularités des fonctions de classe C^m en montrant que l'on pouvait approcher toute fonction numérique f de classe C^m à n variables par des fonctions dont les seuls points singuliers sont des points isolés *critiques* (c'est-à-dire dont la dérivée s'annule) dont la dérivée seconde est une forme bilinéaire associée à une forme quadratique non dégénérée. Il montra en outre qu'il était possible de transmuer un tel point singulier, à l'aide d'un difféomorphisme local, en une fonction qui est une somme algébrique des carrés des coordonnées.

Ainsi, on abandonne l'étude inextricable de *toutes* les singularités possibles pour ne s'intéresser qu'à des singularités *génériques* auxquelles on peut toujours se ramener grâce à une approximation et à une transmutation.

En 1955, H. Whitney résoud la même question concernant les applications du plan $\mathbf{R}^2$ sur le plan $\mathbf{R}^2$. Il met en évidence deux singularités génériques : le *pli* qui se ramène au modèle $(x,y) \mapsto (x^2,y)$ et la *fronce* dont le modèle est $(x,y) \mapsto (x^3 - 3\,xy,\ y)$. Dans ces travaux, un rôle essentiel est joué par le théorème d'A. Sard (1942), dont le cas particulier simple relatif

aux fonctions C^∞ avait déjà été démontré par Morse :

Théorème. Étant donné une application de classe C^m de $\mathbf{R}^n$ dans $\mathbf{R}^p$ la mesure au sens de Lebesgue de l'image des points singuliers est nulle lorsque $m \geqslant n - p + 1$ (Whitney a construit un exemple montrant que cette inégalité est nécessaire).

Utilisant ce résultat, R. Thom a démontré un *théorème de transversalité* qui, généralisant la méthode de Morse, indique dans quelle condition une application f de $\mathbf{R}^n$ dans $\mathbf{R}^p$ peut être approchée par des fonctions n'ayant que des singularités dûment cataloguées.

Les cas $p=2n+1$, $p=2n$ et $p=2n-1$, entièrement élucidés par Whitney, constituent son célèbre *théorème du plongement* selon lequel toute variété différentiable à n dimensions abstraite, peut être réalisée comme sous-ensemble d'un espace $\mathbf{R}^{2n}$. On peut même se contenter d'un espace $\mathbf{R}^{2n-1}$ si l'on accepte de laisser subsister des « self-intersection ». Par exemple, la « bouteille de Klein » peut être réalisée dans $\mathbf{R}^3$ à condition de se traverser elle-même, et dans $\mathbf{R}^4$ sans aucun point double.

Étude des idéaux de fonctions différentiables

À la suite du théorème de H. Whitney, déjà cité, S. Lojasiewicz a démontré en 1959 que tout idéal engendré dans l'algèbre des fonctions C^∞ par une fonction analytique réelle est nécessairement fermé.

Ce théorème signifie qu'une fonction f de classe C^∞ est divisible par une fonction analytique g, si la division « ponctuelle » des séries formelles de Taylor en chaque point du domaine de f et g est possible.

En 1962, B. Malgrange est parvenu à démontrer que le *théorème de préparation* de Weierstrass, classique dans le cas des fonctions analytiques de plusieurs variables (cf. FONCTIONS ANALYTIQUES) reste valable pour les fonctions C^∞.

Il s'agit d'une *division avec reste* (analogue à la division euclidienne) : il est possible de diviser au voisinage de l'origine une fonction C^∞ de $n+1$ variables $(y, X) = (y, x_1, x_2, ..., x_n)$ par un polynôme :

$$y^k + \sum_{i<k} a_i(X) y^i.$$

Ici les a_i sont des fonctions C^∞ qui s'annulent pour $X = 0$, avec un reste qui est un polynôme de degré inférieur à k en y, dont les coefficients sont C^∞ en X.

Dans le cas analytique, on sait qu'une telle division s'effectue de façon unique. Il n'en est plus de même dans le cas C^∞, ce qui fait la difficulté du théorème de Malgrange.

GEORGES GLAESER

Bibliographie

G. BEAUDOIN, *Calcul vectoriel et linéaire*, Presses univ. Laval, 1980 / C. BOUCHET & J.-P. SAGNET, *Calcul différentiel et intégral*, Morin, Chicoutimi (Québec), 1985 / H. CARTAN, *Cours de calcul différentiel*, Hermann, nouv. éd. 1977 / J. DIEUDONNÉ, *Fondements de l'analyse moderne*, Gauthier-Villars, 3e éd. 1979 ; *Calcul infinitésimal*, Hermann, 2e éd. 1980 / J. DOUCHET & B. ZWAHLEN, *Calcul différentiel et intégral*, Presses polytechniques romandes, Lausanne, 1986-1989 / J.-C. TOUGERON, *Idéaux de fonctions différentiables*, Springer, Berlin, 1972 / P. VEREECKE, *Fondements du calcul différentiel*, P.U.F., 1983 ; *Application du calcul différentiel*, *ibid.*, 1985.

CALCUL SYMBOLIQUE

→ SYMBOLIQUE CALCUL

CALCUL TENSORIEL
→ TENSORIEL CALCUL

CALCULS ASYMPTOTIQUES
→ ASYMPTOTIQUES CALCULS

COMBINATOIRE ANALYSE

L'analyse combinatoire est l'ensemble des techniques qui servent, en mathématiques, à *compter* (ou *dénombrer*) certaines structures *finies*, ou à les *énumérer* (établir des listes exhaustives de structures considérées), enfin à démontrer leur *existence* pour certaines valeurs des paramètres dont elles dépendent. Ces structures sont très variées ; leur seul trait commun c'est d'être *finies*. En revanche, les problèmes qu'on se pose sur ces structures sont très divers, et les techniques mathématiques qu'on utilise pour résoudre ces problèmes, très différentes. Par exemple, si on veut dénombrer les arbres de n sommets, on établit une correspondance biunivoque entre l'ensemble de ces arbres et l'ensemble de certaines suites qu'on sait compter. Mais, si l'on veut démontrer l'existence d'une famille infinie de codes correcteurs, on utilise des résultats fins sur les anneaux de polynômes à coefficients dans un corps fini. Pourtant, dans les deux cas, on dit qu'on s'occupe d'analyse combinatoire. Dans le foisonnement des sujets dits de nature combinatoire, on a donc dû faire un choix et laisser de côté des objets importants.

1. Dénombrements élémentaires

Dans les opérations élementaires de dénombrement, on utilise un langage très proche du réel. On parle de *choisir un objet de m façons différentes*, on dit qu'il n'y a qu'un nombre n de possibilités... Considérons ainsi l'exemple suivant. Une urne contient 10 boules numérotées de 1 à 10 ; on tire *successivement* deux boules de l'urne sans remettre la première après tirage. Combien y a-t-il de tirages *croissants*, c'est-à-dire de façons de tirer deux boules dont les numéros vont en croissant ? Pour déterminer ce nombre, on raisonne de la façon suivante. Si la première boule a été tirée et que son numéro est i ($1 \leqslant i \leqslant 10$), pour obtenir un tirage croissant, on peut choisir le numéro j de la seconde de $10 - i$ façons différentes. Enfin, pour obtenir un tirage croissant, on peut choisir soit un tirage commençant par le numéro 1, soit un tirage commençant par 2, etc. Le nombre de tirages croissants est alors égal à $(10 - 1) + (10 - 2) + \ldots + (10 - 9) + (10 - 10) = 45$. On a implicitement utilisé les deux règles suivantes (la seconde avant la première) :

– *Règle de la somme* : Si on peut choisir un objet a de m façons et un objet b de n façons, on peut choisir a ou b de $m + n$ façons.

– *Règle du produit* : Si on peut choisir un objet a de m façons, *puis* un objet b de n façons, on peut choisir *a puis b, dans cet ordre*, de mn façons.

Reprenons l'exemple ci-dessus. Pour caractériser un tirage, il suffit de se donner un couple (i, j) d'entiers tels que i et j soient compris entre 1 et 10. Désignons par X l'ensemble des couples (i, j) tels que

$1 \leqslant i < j \leqslant 10$. Les tirages croissants correspondent aux éléments de X et, pour trouver le nombre de tirages croissants, il suffit de *dénombrer l'ensemble* X, c'est-à-dire de compter le nombre de ses éléments. Pour $1 \leqslant h \leqslant 10$, notons X_h l'ensemble des couples (i, j) de X tels que $i = h$; ainsi X_h est le produit cartésien des ensembles $\{h\}$ et $\{h+1, h+2, ..., 10\}$. De plus, les ensembles X_h sont disjoints deux à deux et leur réunion est X. Désignons par $|X_h|$ le nombre des éléments de X_h ; alors en posant $k = 10$, le nombre, noté $|X|$ des éléments de X est donné par :

(1) $$|X| = |X_1| + |X_2| + ... + |X_k|.$$

De plus, pour $1 \leqslant h \leqslant 10$, on a :

(2) $$|X_h| = |\{h\}| . |\{h+1, h+2, ..., 10\}| = 1 . (10 - h) = 10 - h.$$

En appliquant les formules (1) et (2), on retrouve bien 45 pour le nombre des éléments de X. Dans cette dernière démarche, nous avons résolument adopté le langage de la théorie des ensembles et utilisé certains résultats sur les ensembles *finis*. D'abord comment caractérise-t-on les ensembles finis ? En premier lieu, on définit une *numérotation* d'un ensemble X comme une suite $x_1, x_2, ..., x_n$ d'éléments de X telle $x_i \neq x_j$ pour $i \neq j$ et telle que tout élément de X soit l'un des x_i. Les ensembles finis sont ceux qui possèdent une numérotation. L'entier n qui apparaît dans une numérotation d'un ensemble fini X s'appelle le *cardinal* de X ou le *nombre d'éléments* de X et se note $|X|$. Cet entier ne dépend pas en effet de la numérotation choisie. On convient que $|\varnothing| = 0$. On obtient alors :

– *Formule de la somme* : Si X et Y sont deux ensembles finis, *disjoints*, l'ensemble $X \cup Y$ est fini et l'on a :

(3) $$|X \cup Y| = |X| + |Y|.$$

– *Formule du produit* : Si X et Y sont deux ensembles finis, le produit cartésien $X \times Y$ de X par Y est fini et l'on a :

(4) $$|X \times Y| = |X| . |Y|.$$

Ces deux formules *se démontrent* de la même façon. On détermine une numérotation de $X \cup Y$ et de $X \times Y$, supposant données des numérotations de X et de Y et on vérifie les formules (3) et (4).

Ces formules ont été utilisées, de façon explicite, dans l'exemple, en (1) et (2). Elles sont la traduction dans le langage de la théorie des ensembles des règles de la somme et du produit. On ne parle plus de « choisir un objet a de m façons », mais on considère l'ensemble X des choix possibles pour a et l'on suppose que l'on a $|X| = m$. Cette transcription nous permettra dans la suite d'utiliser indifféremment les deux langages. Lorsque les ensembles X et Y sont des ensembles finis quelconques, on peut exprimer les ensembles $X \cup Y$ et Y comme des réunions de deux ensembles disjoints, à savoir :

$$X \cup Y = X \cup (Y \cap \complement X)$$

et $$Y = (X \cap Y) \cup (Y \cap \complement X).$$

En appliquant la formule de la somme dans les deux cas, on obtient :

(5) $$|X \cup Y| = |X| + |Y| - |X \cap Y|.$$

Rappelons enfin un principe fondamental dans le dénombrement :

(6) Pour que deux ensembles finis X et Y aient le même cardinal, il faut et il suffit qu'il existe une bijection de X sur Y.

Dans de nombreux cas, la difficulté sera effectivement de *construire* une telle bijection. Examinons maintenant quelques dénombrements fondamentaux.

D'abord les trois formules (3), (4) et (5) s'étendent au cas général de k ensembles ($k \geqslant 2$). On obtient :

$$(7) \quad |X_1 \cup X_2 \cup \ldots \cup X_k| = |X_1| + |X_2| + \ldots + |X_k|,$$

si $X_i \cap X_j = \emptyset$, pour $i \neq j$.

De même, on a toujours :

$$(8) \quad |X_1 \times X_2 \times \ldots \times X_k| = |X_1| . |X_2| \ldots |X_k|.$$

Quant à la formule (5), sa généralisation est la suivante :

$$(9) \quad |X_1 \cup X_2 \cup \ldots \cup X_k| = \sum_i |X_i| - \sum_{i_1 \leqslant i_2} |X_{i_1} \cap X_{i_2}| + \ldots + (-1)^{r-1} \sum_{i_1 < i_2 < \ldots < i_r} |X_{i_1} \cap X_{i_2} \cap \ldots \cap X_{i_r}| + \ldots + (-1)^{k-1} |X_1 \cap X_2 \cap \ldots \cap X_k|.$$

La formule (9) peut s'établir aisément par récurrence sur k. On l'appelle la formule du *principe d'inclusion et d'exclusion.* Elle a reçu de nombreuses applications.

Si, dans la formule (8), on prend tous les ensembles X_i égaux au même ensemble X, on a alors :

$$(10) \qquad |X^k| = |X|^k.$$

Ainsi, l'ensemble X^k de tous les k-uples $(x_1, x_2, \ldots, x_k)$, où les x_i sont pris dans X, a pour cardinal $|X|^k$. Dans la littérature classique, si $|X| = n$, un k-uple $(x_1, x_2, \ldots, x_k)$ était appelé *arrangement avec répétition de n objets pris k à k.*

Soit maintenant $y_1, y_2, \ldots, y_k$ une numérotation d'un ensemble fini Y. Pour définir une application f de Y dans X, il suffit de se donner arbitrairement des éléments $x_1, x_2, \ldots, x_k$ (non nécessairement distincts) de X et de poser $f(y_i) = x_i$ pour $1 \leqslant i \leqslant k$. On définit ainsi une bijection de l'ensemble X^k des k-uples $(x_1, x_2, \ldots, x_k)$ pris dans X sur l'ensemble, noté X^Y, de toutes les applications de Y dans X. La formule (10) et le principe (6) impliquent donc :

$$(11) \qquad |X^Y| = |X|^{|Y|}.$$

D'autre part, l'application qui fait correspondre à toute partie A de X sa fonction caractéristique φ_A est une bijection de l'ensemble, noté $\mathfrak{P}(X)$, de toutes les parties de X sur l'ensemble $\{0, 1\}^X$ des applications de X dans $\{0, 1\}$. La formule précédente implique alors $|\mathfrak{P}(X)| = |\{0, 1\}|^{|X|} = 2^{|X|}$. Soit

$$(12) \qquad |\mathfrak{P}(X)| = 2^{|X|}.$$

Dénombrement des injections

Considérons deux ensembles finis X et Y avec $|X| = n$, $|Y| = p$ et $n \leqslant p$. Soit $\mathfrak{I}$ l'ensemble des *injections* de X dans Y, c'est-à-dire des applications f de X dans Y telles que les relations $x, x' \in X$, $x \neq x'$ entraînent $f(x) \neq f(x')$. Il est clair que $|\mathfrak{I}|$ ne dépend que de p et de n ; posons $|\mathfrak{I}| = A(p, n)$. Prenons un ensemble X′ disjoint de X avec $|X'| = p - n$. L'ensemble $X'' = X \cup X'$ a ainsi p éléments. Pour définir une bijection de X″ sur Y, il suffit de se donner une injection f de X dans Y et une bijection g de X′ sur $Y - f(X)$. On peut choisir f de $A(p, n)$ façons, puis g de $A(p - n, p - n)$ façons. La règle du produit permet ainsi d'écrire : $A(p, p) = A(p, n)\, A(p - n, \ p - n)$, où $0 \leqslant n \leqslant p$. Comme on a : $A(p, 1) = p$ et $A(p, 0) = 1$ (en convenant qu'il y a une application d'un ensemble vide dans tout ensemble), il vient : $A(p, p) = pA(p - 1, p - 1)$. D'où $A(p, p) = p\,!$, en posant

$0\,! = 1$ et $p\,! = p((p-1)\,!)$ pour $p \geqslant 1$. On a enfin :

(13) $A(p, n) = p\,!/(p-n)!$
$= p(p-1) \dots (p-n+1).$

Soit $x_1, x_2, \dots, x_n$ une numérotation donnée de X ; pour définir une injection f de X dans Y, il suffit encore de se donner arbitrairement une suite de n éléments $(y_1, y_2, \dots, y_n)$ de Y *tous distincts* et de poser $f(x_i) = y_i$ pour $1 \leqslant i \leqslant n$. On trouve ainsi que le nombre de n-uples $(y_1, y_2, \dots, y_n)$, où les y_i sont des éléments tous distincts, pris dans un ensemble Y de cardinal p, est égal à $p\,!/(p-n)\,!$. Dans les ouvrages classiques, un tel n-uple est appelé *arrangement sans répétition de p objets pris n à n*.

Si on prend Y = X, la formule (13) implique que le nombre de bijections de X sur lui-même – on dit aussi *permutations* de X – est égal à $n\,!$.

Soit C_p^n le nombre de parties de Y qui ont n éléments. Pour définir une injection de X dans Y, il suffit de se donner une partie A de Y ayant n éléments, puis une bijection f de X sur A. On peut choisir A de C_p^n façons, puis f de n ! façons. On a donc $|\mathfrak{J}| = C_p^n\, n\,!$. D'où $C_p^n = A(p, n)/n\,! = p\,!/(n\,!(p-n)\,!)$. Les nombres C_p^n sont les *coefficients binomiaux* : C_p^n est aussi noté fréquemment : $\binom{p}{n}$. Ce sont eux qui apparaissent dans la *formule* dite *du binôme* :

$$(x+y)^p = \sum_{n=0}^{p} C_p^n x^n y^{p-n},$$

valable dans tout anneau commutatif. Notons que l'on a : $C_p^0 = C_p^p = 1$, $C_p^1 = p$, $C_p^n = C_p^{p-n}$. On vérifie également la propriété $C_p^n = C_{p-1}^n + C_{p-1}^{n-1}$, qui permet de les déterminer de proche en proche. Le coefficient C_p^n est aussi le nombre des *combinaisons de p éléments pris n à n*, mais on tend à abandonner cette terminologie, une combinaison n'étant qu'un sous-ensemble A, de cardinal n, d'un ensemble Y de cardinal p.

Dénombrement des applications croissantes de X dans Y

Supposons que l'on prenne les ensembles $X = \{1, 2, \dots, n\}$ et $Y = \{1, 2, \dots, p\}$, où les entiers n et p sont quelconques. Une application f de X dans Y est dite *croissante* si l'on a $f(i) \leqslant f(j)$ pour tout couple (i, j) tel que $1 \leqslant i \leqslant j \leqslant n$. Par un raisonnement déjà utilisé plus haut, on vérifie que l'ensemble Γ de toutes les applications croissantes de X dans Y a même cardinal que l'ensemble Γ' des suites *croissantes* $(y_1, y_2, \dots, y_n)$ de n termes pris dans Y, c'est-à-dire qui satisfont à $y_1 \leqslant y_2 \leqslant \dots \leqslant y_n$. Une telle suite est encore appelée *combinaison avec répétition de p objets pris n à n*. À son tour, Γ' a même cardinal que l'ensemble Γ'' des suites $(z_1, z_2, \dots, z_n)$ de n termes où les z_i sont pris dans l'ensemble $\{1, 2, \dots, p+n-1\}$ et satisfont à $z_1 < z_2 < \dots < z_n$. On établit en effet une bijection de Γ' sur Γ'' en faisant correspondre à la suite $(y_1, y_2, \dots, y_n)$ de Γ' la suite $(y_1 + 0, y_2 + 1, \dots, y_n + n - 1)$. On a ainsi $|\Gamma| = |\Gamma''| = C_{p+n-1}^n$.

Dénombrement des surjections

Soit X et Y deux ensembles finis, respectivement de cardinal n et p. Désignons par $\mathfrak{S}$ l'ensemble des *surjections* de X sur Y, c'est-à-dire des applications f de X dans Y, telles que pour tout $y \in Y$, il existe un $x \in X$ satisfaisant à $f(x) = y$. L'ensemble $\mathfrak{S}$ est vide si $n < p$. Supposons désormais que l'on a $n \geqslant p$. On ne connaît pas de formule explicite donnant le cardinal de $\mathfrak{S}$ en fonction de n et p. Toutefois, il existe une formule de récurrence, qui permet de

le calculer de proche en proche. On appelle d'abord *partition* de l'ensemble X la donnée de r sous-ensembles $A_1, A_2, ..., A_r$ ($r > 0$) de X (dont certains peuvent être vides) satisfaisant à $A_i \cap A_j = \emptyset$ si $i \neq j$ et $X = A_1 \cup A_2 \cup ... \cup A_r$. Notons que pour définir une partition, on ne s'impose pas un ordre sur les A_i. Désignons par S_n^p *le nombre de partitions de* X *en p sous-ensembles non vides.* Pour définir ensuite une surjection f de X sur Y, il suffit de se donner une partition de X en p sous-ensembles non vides, puis une bijection de l'ensemble formé par ces p sous-ensembles, sur Y. D'après la règle du produit, on a $|\mathcal{S}| = p\,!\, S_n^p$. Il est maintenant facile de vérifier que les nombres S_n^p satisfont aux relations de récurrence suivantes :

$$S_n^1 = S_n^n = 1 \text{ et } S_{n+1}^p = S_n^{p-1} + p S_n^p$$

pour $1 < p < n$.

Ces relations de récurrence permettent donc de calculer $|\mathcal{S}|$ de proche en proche. Les coefficients S_n^p s'appellent les *nombres de Stirling* (de deuxième espèce). Donnons le tableau des premiers coefficients S_n^p :

	$p =$ 1	2	3	4	5
$n =$ 1	1	0	0	0	0
2	1	1	0	0	0
3	1	3	1	0	0
4	1	7	6	1	0
5	1	15	25	10	1

Par exemple, le nombre de surjections d'un ensemble de 5 éléments sur un ensemble de 3 éléments est égal à $3\,!\, S_5^3 = 150$.

2. Fonctions génératrices

On a vu qu'on ne savait pas trouver de formule explicite donnant le nombre de surjections d'un ensemble de n éléments sur un ensemble de p éléments, mais on peut faire apparaître ces nombres comme coefficients d'une série. La fonction de deux variables $f(t, u) = \exp(t(\exp u - 1))$ se développe en série sous la forme :

$$1 + \sum_{n \geqslant 1} \frac{u^n}{n\,!} \cdot \sum_{p=1}^{n} S_n^p t^p ;$$

le nombre de Stirling S_n^p apparaît, divisé par $n\,!$, comme coefficient du monôme $u^n t^p$. On dit que la fonction $f(t, u)$ est la *fonction génératrice* des nombres $S_n^p/n\,!$; puisqu'on connaît l'expression de cette fonction génératrice, on pourra résoudre d'autres problèmes combinatoires ou trouver des fonctions génératrices d'autres nombres liés aux nombres de Stirling.

Supposons donné un anneau A commutatif et ayant un élément unité. On appelle *série formelle* à coefficients dans A et en une indéterminée u, une somme symbolique infinie :

$$\sum_{n \geqslant 0} a_n u^n ,$$

où les a_n sont dans A ($n \geqslant 0$). On dit que a_n est le *coefficient* de u^n dans cette série. La somme et le produit de deux séries formelles sont définis de manière à respecter les règles usuelles de distributivité pour le produit de deux séries infinies et le calcul sur les puissances $u^n u^m = u^{n+m}$ (cf. ANNEAUX ET ALGÈBRES, chap. 2) : si on a deux séries :

$$\sum_{n \geqslant 0} a_n u^n \quad \text{et} \quad \sum_{n \geqslant 0} b_n u^n ,$$

leur somme est la série :

$$\sum_{n \geqslant 0} (a_n + b_n) u^n ,$$

et leur produit est donné par :

$$\sum_{n \geqslant 0} c_n u^n,$$

où l'on a : $c_n = a_0 b_n + a_1 b_{n-1} + ... + a_n b_0$ pour tout $n \geqslant 0$. Si U est une telle série sans terme constant, c'est-à-dire telle que le coefficient de u^0 soit nul, on appelle exponentielle de U la série :

$$\exp \mathrm{U} = \sum_{n \geqslant 0} \mathrm{U}^n/n\,!.$$

En d'autres termes, exp U est la série formelle dont le coefficient de u^n est le coefficient de u^n dans la somme (finie) :

$$1 + \mathrm{U}/1! + \mathrm{U}^2/2! + ... + \mathrm{U}^n/n\,!\,(n \geqslant 0).$$

Une suite infinie d'indéterminées $\mathbf{t} = (t_1, t_2, ...)$ étant donnée, nous prenons pour anneau A l'anneau des polynômes à coefficients rationnels dont les indéterminées sont prises dans la suite **t**. Posons alors :

$$\mathrm{U} = \sum_{n \geqslant 1} t_n u^n/n\,!\,;$$

son terme constant est nul, on peut donc prendre son exponentielle, qu'on peut écrire :

$$\exp \mathrm{U} = 1 + \sum_{n \geqslant 1} a_n u^n/n\,!.$$

Il est facile de voir que a_n est un polynôme de degré n en les n variables t_1, t_2, ..., t_n, où $n \geqslant 1$. Plus précisément, on a :

$$(1) \quad a_n = a_n(t_1, t_2, ..., t_n) = \sum (n\,!/(m_1!\,m_2!\,...\,m_n!)) \times (t_1/1!)^{m_1}(t_2/2!)^{m_2}...(t_n/n\,!)^{m_n},$$

où la sommation est étendue à toutes les suites $(m_1, m_2, ..., m_n)$ de n entiers positifs satisfaisant à $1m_1 + 2m_2 + ... + nm_n = n$.

Le coefficient du monôme :

$$t_1^{m_1} t_2^{m_2} ... t_n^{m_n}$$

dans a est donc égal à :

$$\mathrm{C}(m_1, m_2 ..., m_n) = n\,!/(m_1!\,m_2!\,...\,m_n!\,1\,!^{m_1}\,2\,!^{m_2}\,...\,n\,!^{m_n}).$$

On peut vérifier que ce nombre est le *nombre de partitions de type* $(m_1, m_2, ..., m_n)$, c'est-à-dire de partitions de l'ensemble $\mathrm{X} = \{1, 2, ..., n\}$ en $m_1 + m_2 + ... + m_n$ sous-ensembles non vides, parmi lesquels m_1 sont de cardinal 1, m_2 de cardinal 2, ..., m_n de cardinal n.

Cette interprétation nous permet de retrouver le nombre de partitions de l'ensemble $\mathrm{X} = \{1, 2, ..., n\}$ en p sous-ensembles non vides. En effet, le nombre de sous-ensembles non vides d'une partition de type $(m_1, m_2, ..., m_n)$ est $p = m_1 + m_2 + ... + m_n$. Si donc l'on pose $t_1 = t_2 = ... = t_n = t$ dans l'expression (1), le coefficient de t^p dans le second membre va être égal au nombre de partitions de X en p sous-ensembles non vides, soit :

$$a_n(t, t, ..., t) = \sum_{p=1}^{n} \mathrm{S}_n^p t^p.$$

Par conséquent la fonction génératrice des nombres de Stirling de seconde espèce est donnée par :

$$(2) \quad 1 + \sum_{n \geqslant 1} (u^n/n\,!) \sum_{p=1}^{n} \mathrm{S}_n^p t^p. = \exp(t(\exp u - 1)).$$

Par exemple, pour trouver le nombre de surjections d'un ensemble de n éléments sur un ensemble de p éléments ($p \leqslant n$), on détermine le coefficient c de $u^n t^p$ dans le développement de exp $(t(\exp u - 1))$ et le nombre cherché est égal à $n\,!\,p\,!\,c$.

3. Construction de correspondances

Dans la première partie, nous avons passé en revue tous les ensembles finis qu'il était aisé de dénombrer en faisant usage des deux règles de la somme et du produit. Ces techniques élémentaires s'appliquent plus difficilement lorsqu'on veut dénombrer d'autres structures finies plus élaborées comme celles des *arbres* ou certains types de *graphes*. Le plus souvent on est conduit à chercher une correspondance biunivoque entre ces structures et les ensembles finis considérés dans la première partie. Jusqu'ici, il n'existe pas de théorie pour construire ces correspondances, tout est une question d'ingéniosité et de patience. À titre d'exemple, on peut décrire ci-dessous une telle correspondance entre les arbres de n sommets et les $(n-2)$-uples $(x_1, x_2, ..., x_{n-2})$, où les x_i sont des entiers compris entre 1 et n.

Un *graphe* est la donnée d'un ensemble X – ses éléments sont les *sommets* du graphe – et d'une classe U de sous-ensembles de X à deux éléments, appelés *arêtes*. Si l'on a $x, y \in X$ et $\{x, y\} \in U$, on dit que x et y sont *adjacents*. Une *chaîne* du graphe $\{X, U\}$ est une suite $\{x_1, y_1\}$, $\{x_2, y_2\}$, ..., $\{x_n, y_n\}$ d'arêtes distinctes telle que, lorsque $n > 1$, on ait :

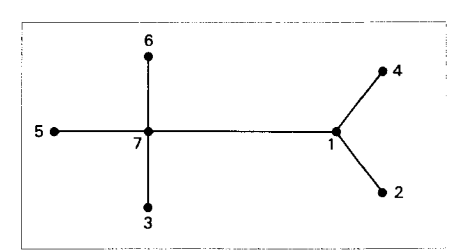
$$\{x_i, y_i\} \cap \{x_{i+1}, y_{i+1}\} \neq \emptyset,$$

pour $i = 1, 2, ..., n-1$. Si de plus, on a $n > 1$ et

$$\{x_n, y_n\} \cap \{x_1, y_1\} \neq \emptyset,$$

on dit que la chaîne est un *cycle*. Un graphe est dit *connexe* si, pour tout couple de sommets distincts (x, y), il existe une chaîne $\{x_1, y_1\}$, $\{x_2, y_2\}$, ..., $\{x_n, y_n\}$ telle que $x_1 = x$ et $y_n = y$. Un *arbre* de n sommets est alors défini comme un graphe de n sommets qui est connexe et sans cycles. On a l'habitude de représenter un graphe fini de n sommets comme un ensemble de n points du plan numérotés de 1 à n où les sommets i et j sont reliés entre eux si et seulement si l'on a $\{i, j\} \in U$. Considérons par exemple le graphe de la figure ci-dessous. C'est un arbre de sept sommets, numérotés de 1 à 7.

Soit A_n l'ensemble de tous les arbres possibles de n sommets numérotés 1, 2, ..., n. Cayley fut le premier à démontrer que l'on a $|A_n| = n^{n-2}$. Or ce nombre est précisément le cardinal de l'ensemble H_n de tous les $(n-2)$-uples $(x_1, x_2, ..., x_{n-2})$ où les x_i sont pris dans $X = \{1, 2, ..., n\}$ (formule (11) de la première partie). Pour démontrer la formule $|A_n| = n^{n-2}$, il suffit donc de construire une bijection Φ de A_n sur H_n. Nous donnons ici la construction d'une telle correspondance due à Prüfer. Celle-ci fait usage des deux propriétés suivantes sur les arbres, faciles à vérifier. D'abord, un arbre a toujours (au moins) un sommet *pendant*, c'est-à-dire un sommet qui n'est adjacent qu'à un seul autre sommet.



Ensuite, si l'on « efface » un sommet pendant d'un arbre de n sommets, et l'arête qui le contient, on obtient un arbre de $(n-1)$ sommets. La construction de Prüfer est alors la suivante : pour un arbre de A_n donné, on efface *le plus petit* sommet pendant et l'on désigne par x_1 l'unique sommet qui était adjacent à ce sommet pendant. On répète cette opération avec l'arbre restant

de $(n-1)$ sommets et l'on détermine x_2 et ainsi de suite. On s'arrête lorsqu'il ne reste plus que deux sommets (adjacents). Par exemple, à l'arbre de la figure ci-dessus on fait correspondre la suite (1, 7, 1, 7, 7). Les sommets pendants qu'on a successivement effacés sont 2, 3, 4, 1, 5 et il est resté l'arbre formé par les deux sommets 6 et 7. On vérifie que l'application Φ de A_n sur H_n ainsi construite est bien bijective. D'où l'on déduit $|A_n| = |H_n| = n^{n-2}$.

4. Théorèmes d'existence

Le théorème de Ramsey et celui de Hall-König, dont nous donnons les énoncés maintenant, jouent un rôle spécial en combinatoire. Ils assurent l'existence de certaines configurations dans des conditions très générales et ont ainsi trouvé de nombreuses applications.

Supposons que l'on ait un graphe de six sommets, dans lequel deux sommets distincts sont joints par une arête « coloriée » en bleu ou en rouge. Démontrons la propriété (P) suivante : Il existe un triangle dont les trois arêtes ont la même couleur. Considérons en effet les cinq arêtes issues d'un sommet donné S_1 du graphe. Nécessairement trois de ces arêtes ont la même couleur, disons bleue ; appelons-les S_1S_2, S_1S_3 et S_1S_4. Si l'une des arêtes S_2S_3, S_2S_4, S_3S_4 est bleue, disons S_2S_3, le triangle $S_1S_2S_3$ est bleu. Sinon les arêtes S_2S_3, S_2S_4, S_3S_4 sont rouges, mais alors le triangle $S_2S_3S_4$ est rouge. La propriété (P) est ainsi vérifiée. L'argument essentiel qui a été utilisé est le *principe des tiroirs* : si l'on met n objets dans p tiroirs ($p \leqslant n$), alors le nombre des objets dans au moins un tiroir est supérieur ou égal à n/p. Le théorème de Ramsey peut être considéré comme une généralisation de ce principe.

Théorème de Ramsey. Soit X un ensemble de n éléments et supposons donnés d'abord trois entiers p, q, r satisfaisant à $p \geqslant r$, $q \geqslant r$ et $r \geqslant 1$, ensuite une partition de l'ensemble de toutes les parties de X de cardinal r, en deux classes $\mathfrak{P}$ et $\mathfrak{Q}$. Alors il existe un entier $N(p, q, r)$ ne dépendant que des entiers p, q, r et non pas de l'ensemble X tel que la propriété (P′) suivante soit vérifiée : Si $n \geqslant N(p, q, r)$, il existe ou bien une partie A de X de p éléments qui a tous ses sous-ensembles de cardinal r dans $\mathfrak{P}$, ou bien une partie B de q éléments qui a tous ses sous-ensembles de cardinal r dans $\mathfrak{Q}$.

Dans l'exemple précédent, on avait $n = 6, p = q = 3$ et $r = 2$. Les ensembles de deux éléments sont alors les arêtes, qui sont divisées en deux classes, les bleues (classe $\mathfrak{P}$) et les rouges (classe $\mathfrak{Q}$). La propriété (P) équivaut à dire que l'on a $N(3, 3, 2) \geqslant 6$. En fait, on a l'égalité.

Le théorème de Hall-König peut être énoncé de deux façons différentes, soit en termes de systèmes de représentants distincts, soit en termes de matrices à coefficients 0 ou 1. Nous ne donnerons pas ici l'énoncé sous cette dernière forme. Soit $X = \{1, 2, 3, 4, 5\}$ l'ensemble des cinq premiers entiers et considérons les sous-ensembles $X_1 = \{1, 4\}$, $X_2 = \{1, 2, 4\}$, $X_3 = \{1, 2, 4\}$ et $X_4 = \{2, 4, 5\}$ de X. On peut choisir des entiers x_1, x_2, x_3, x_4 de sorte que l'on ait $x_i \in X_i$ pour $1 \leqslant i \leqslant 4$ et que tous les x_i soient *distincts*. Il suffit de prendre en effet les nombres $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 5$. Si on remplace dans cet exemple X_4 par $X_4' = \{2, 4\}$, un tel choix n'est plus possible puisque la réunion des quatre ensembles X_1, X_2, X_3, X_4' ne contient que trois éléments distincts 1, 2, 4. On est ainsi amené à donner la définition suivante : Étant donné une suite $(X_1, X_2, ..., X_n)$ de parties

d'un ensemble non vide X, on dit que cette suite possède un *système de représentants distincts*, s'il existe une suite $(x_1, x_2, ..., x_n)$ de n éléments *distincts* de X satisfaisant à $x_i \in X_i$ pour tout i tel que $1 \leqslant i \leqslant n$.

Théorème de Hall-König. La condition nécessaire et suffisante pour qu'une suite $(X_1, X_2, ..., X_n)$ de parties d'un ensemble non vide X possède un système de représentants distincts est que, pour tout k $(1 \leqslant k \leqslant n)$ et tout sous-ensemble $\{i_1, i_2, ..., i_k\}$ de k indices extrait de $\{1, 2, ..., n\}$ on ait $|X_{i_1} \cup X_{i_2} \cup ... \cup X_{i_k}| \geqslant k$.

La condition nécessaire est tout à fait évidente, mais la condition suffisante est loin de l'être.

5. Existence et construction de modèles

Un certain nombre de modèles ont été tout particulièrement étudiés, c'est le cas des *carrés latins*, sans doute parce qu'un mathématicien célèbre comme Euler fit à leur sujet une conjecture malheureuse et qu'il fallut attendre 177 ans pour prouver son inexactitude. En introduisant des notions comme celle d'orthogonalité, on a pu établir des liens étroits entre les carrés latins et certaines géométries finies, ou encore avec d'autres modèles comme les *blocs incomplets équilibrés*, ce qui a permis souvent d'étudier le même objet avec des optiques différentes.

Un *carré latin* d'ordre n est une matrice carrée $A = (a_{ij})$, $_{1 \leqslant i, j \leqslant n}$ dont les coefficients a_{ij} sont $1, 2, ..., n$ et dans laquelle chaque entier k apparaît une et une seule fois dans chaque ligne et chaque colonne $(1 \leqslant k \leqslant n)$. Il existe naturellement un carré latin d'ordre n pour tout $n \geqslant 1$. Par exemple, la table de multiplication d'un groupe fini d'ordre n est un carré latin. En fait, un carré latin peut être considéré comme la table de multiplication d'un système algébrique dont la loi est seulement supposée simplifiable. Soit l_n le nombre de carrés latins d'ordre n ; jusqu'à ce jour, on a pu déterminer les valeurs exactes de l_n pour $1 \leqslant n \leqslant 8$. Pour calculer l_8, on a dû recourir à l'usage d'ordinateurs, mais même avec ceux-ci, en l'absence de nouvelles méthodes, on ne peut espérer aller bien loin dans cette direction. En revanche, l'orthogonalité a permis de trouver des résultats plus intéressants. Soit $A = (a_{ij})$ et $B = (b_{ij})$, deux carrés latins d'ordre n ; on dit qu'ils sont *orthogonaux* si tous les n^2 couples (a_{ij}, b_{ij}), où $i, j = 1, 2, ..., n$, sont distincts. Par exemple, les deux carrés latins d'ordre 4 indiqués ci-dessous sont orthogonaux :

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

Soit C la matrice ayant pour coefficients les couples (a_{ij}, b_{ij}) ; dans C tous les n^2 couples $(1, 1), (1, 2), ..., (n, n)$ apparaissent une et une seule fois. Dans l'exemple ci-dessus, on obtient la matrice :

$$C = \begin{pmatrix} 1,1 & 2,2 & 3,3 & 4,4 \\ 2,3 & 1,4 & 4,1 & 3,2 \\ 3,4 & 4,3 & 1,2 & 2,1 \\ 4,2 & 3,1 & 2,4 & 1,3 \end{pmatrix}$$

On dit parfois que C est un *carré gréco-latin*. On s'est alors posé le problème de savoir s'il existe une paire de carrés latins orthogonaux pour tout n. Avec un peu de patience, il est facile de construire de telles paires pour $n = 3, 4$ et 5 et même pour des entiers supérieurs à 6. C'est Euler qui en 1782 formula ce problème en termes

récréatifs. Supposons, en effet, que, dans six régiments donnés, on relève, dans chacun, six officiers, tous de grades différents, et qu'on veuille disposer ces trente-six officiers dans un carré de six cases de côté, de sorte que dans chaque ligne et dans chaque colonne il y ait un représentant de chaque régiment et un représentant de chaque grade. Une telle disposition est-elle possible ? La réponse est non, et ce n'est qu'en 1900 que Tarry démontra cette impossibilité par une analyse très systématique de tous les cas possibles. Le *problème des trente-six officiers* était donc impossible. Si l'on numérote de 1 à 6 les six régiments et les six grades, chacun des officiers peut être repéré par un couple (i, j), où $1 \leqslant i, j \leqslant 6$; le premier indice i désigne son régiment et le second j son grade. Vouloir un représentant de chaque régiment (resp. grade) dans chaque ligne et chaque colonne et vouloir quand même disposer les trente-six officiers, c'est chercher à satisfaire aux trois exigences :

- Les premiers indices i forment un carré latin ;
- Les seconds indices j forment un carré latin ;
- Les deux carrés latins ainsi formés sont orthogonaux.

L'impossibilité du problème des trente-six officiers démontrée par Tarry équivaut donc à l'impossibilité de construire deux carrés latins d'ordre 6 orthogonaux. Euler n'ayant pu réussir à construire de couples de carrés latins orthogonaux d'ordre n pour des entiers de la forme $n = 4k + 2$, conjectura qu'il n'existait pas de couples de carrés latins orthogonaux d'ordre n pour tout n de la forme $4k + 2$. À l'exception du seul cas $n = 6$, sa conjecture s'est révélée complètement erronée et, en 1959, Bose, Shrikhande et Parker démontrèrent qu'il existe en effet un couple de carrés latins orthogonaux d'ordre n pour tout n différent de 2 et 6. Avant cette date, cependant, on savait construire un couple de carrés latins orthogonaux pour tout n non congru à 2 modulo 4. Une des méthodes qui s'est révélée des plus fécondes a été de considérer pour tout $n \geqslant 3$, de la forme p^r, où p est un nombre premier et $r > 0$, le corps fini F_n ayant n éléments. Désignons par $x_o = 0$, x_1, ..., x_{n-1} les n éléments de F_n et formons les matrices A_1, A_2, ..., A_{n-1}, où $A_k = ({}_k a_{ij})$ avec ${}_k a_{ij} = x_k x_i + x_j$ $(i, j = 0, 1, ..., n-1$; $k = 1, 2, ..., n-1)$. Il est facile de vérifier que A_1, A_2, ..., A_{n-1} sont des carrés latins orthogonaux deux à deux. On a pu étendre cette construction à tous les entiers $n \not\equiv 2$ modulo 4 et démontrer que si $p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}$ est la décomposition d'un entier n en facteurs premiers (les p_i étant tous distincts et les r_i strictement positifs) et si $t = \min (p_i^{r_i} - 1)$ pour $1 \leqslant i \leqslant k$ est supérieur ou égal à 2, alors il existe t carrés latins d'ordre n orthogonaux deux à deux.

Notons que pour $n \geqslant 3$, il est aisé de démontrer que le cardinal de tout ensemble de carrés latins orthogonaux deux à deux est au plus égal à $(n - 1)$. Lorsque n est supérieur ou égal à 3 et de la forme p^r où p est premier et $r > 0$, d'après ce qui précède, on peut construire un ensemble de carrés latins orthogonaux deux à deux qui soit de cardinal maxima, à savoir $(n - 1)$. Qu'en est-il pour les autres entiers ? La question est loin d'être résolue. Le théorème de Bruck-Ryser démontré en 1949 et énoncé ci-dessous apporte un premier élément de réponse. Il affirme la non-existence de certaines configurations dites *plans projectifs d'ordre n*, pour une famille infinie d'entiers. Sans vouloir donner de nouvelles définitions, disons que l'existence d'un plan projectif d'ordre n est équivalente à l'existence d'un ensemble de

$(n-1)$ carrés latins d'ordre n orthogonaux deux à deux.

Théorème de Bruck-Ryser. Si pour $n \equiv 1$ ou 2 mod 4, il existe un nombre premier p de la forme $4k+3$ et un entier $r > 0$ tels que p^{2r-1} divise n et p^{2r} ne divise pas n, alors il n'existe pas de plan projectif d'ordre n.

Il n'y a, par exemple, pas de plan projectif d'ordre 14 et par conséquent on *ne* peut trouver 13 carrés latins d'ordre 14 orthogonaux deux à deux.

Donnons pour terminer un bref aperçu sur les modèles dits en *blocs incomplets équilibrés.* Soit X un ensemble de v éléments $x_1, x_2, ..., x_v$ et $X_1, X_2, ..., X_b$ des sous-ensembles, dits blocs, de X qui soient distincts. On dit que ces sous-ensembles forment un *modèle en blocs incomplets équilibrés* de paramètres (b, v, r, k, λ), si les propriétés suivantes sont vérifiées :

(I) $|X_i| = k$ pour $1 \leqslant i \leqslant b$;

(II) tout $x_i \in X$ appartient à exactement r blocs $(1 \leqslant i \leqslant v)$;

(III) toute paire d'éléments de X apparaît dans λ blocs.

Les paramètres b, v, r, k et λ ne sont pas indépendants. Il est aisé de vérifier les deux relations :

(1) $bk = vr$ et $r(k-1) = \lambda(v-1)$.

Mais à leur tour, si b, v, r, k et λ satisfont aux relations (1), cela n'implique pas nécessairement qu'il existe un modèle en blocs incomplets équilibrés de paramètres (b, v, r, k, λ). Beaucoup de résultats partiels sont connus au sujet de la construction de tels modèles. Lorsque $b = v$ (donc $k = r$), on dit que le modèle est *symétrique* ; on l'appelle encore (v, k, λ)-*configuration* et l'on montre qu'un plan projectif d'ordre n est équivalent à une (v, k, λ)-configuration de paramètres $v = n^2 + n + 1$, $k = n + 1$ et $\lambda = 1$.

DOMINIQUE FOATA

Bibliographie

R. A. Brualdi, *Introductory Combinatorics*, Elsevier Science Publ., 1991 / R. C. Bose & B. Manvel, *Introduction to Combinatorial Theory*, Wiley, New York, 1984 / L. Comtet, *Analyse combinatoire*, 2 vol., P.U.F., Paris, 1970 / O. Favaron, M. Maheo & I. Fournier, *Combinatoire et algorithmique*, Orsay-Plus, 1990 / D. Foata dir., *Combinatoire et représentation du groupe symétrique*, Springer-Verlag, Berlin, 1977 / M. Hall Jr., *Combinatorial Theory*, Wiley, New York, 1986 / P. Raymond, *De la combinatoire aux probabilités : la combinatoire de Cardan à Jacques Bernoulli*, Maspero, 1975 / J. Riordan, *Introduction to Combinatorial Analysis*, Princeton Univ. Press, Princeton, 1980 / A. Slomson, *An Introduction to Combinatorics*, Chapman & Hall, 1991 / W. A. Whitworth, *Choice and Chance, with One Thousand Exercices*, Londres, 1901, réimpr. Hafner Publ., New York, 1965.

COMPLEXES NOMBRES

Introduits à l'origine comme symboles purement formels destinés à rendre compte des propriétés des équations algébriques, les nombres imaginaires sont d'un usage courant au XVIIIe siècle, mais ce n'est qu'au siècle suivant qu'ils seront définis et utilisés correctement, avec la rigueur qui caractérise les préoccupations des mathématiciens du XIXe siècle. Et c'est alors le prodigieux essor de la théorie des fonctions d'une variable complexe et l'entrée en force des imaginaires dans presque tous les domaines des mathématiques. De nos jours, les nombres complexes interviennent de manière essentielle, comme un cadre naturel, dans maintes théories mathématiques et physiques.

1. Historique

Les nombres « impossibles »

Alors que de nombreux mathématiciens (dont Viète) hésitaient encore à utiliser les nombres négatifs, les algébristes italiens du XVIe siècle, Cardan et ses élèves, s'enhardirent à introduire dans les calculs des symboles purement formels $\sqrt{-a}$, $a > 0$, représentant le résultat de l'extraction « impossible » de la racine carrée du nombre négatif $-a$; ils décrivent en détail des règles de calcul permettant de manipuler ces nouveaux « nombres », appelés par eux *nombres impossibles.*

À l'origine, il s'agissait seulement de donner des racines à *toutes* les équations du second degré ; les résultats obtenus dans l'étude de l'équation du troisième degré allaient familiariser les mathématiciens avec ces symboles et mettre en évidence leur rôle comme intermédiaire commode de calcul dans de nombreux cas. Au moyen de la formule dite de Cardan, Bombelli montre, en 1572, que la racine $x = 4$ de l'équation $x^3 = 15x + 4$ peut s'écrire :

$$\sqrt[3]{2-\sqrt{-121}} + \sqrt[3]{2+\sqrt{-121}} = 4,$$

mettant par là en évidence le fait que certaines quantités réelles peuvent être représentées par des expressions en apparence imaginaires. Ainsi, la formule de Cardan permet de représenter des racines réelles par l'intermédiaire d'opérations effectuées sur des nombres impossibles, ou « imaginaires ». Les nombres imaginaires fournissent donc des méthodes de calcul, de nature certes mystérieuse, mais qui permettent d'obtenir des résultats « vrais » qu'il serait souvent beaucoup plus long ou beaucoup plus difficile d'obtenir directement.

Pour ces raisons, somme toute empiriques, les mathématiciens utilisèrent avec une confiance croissante les nombres imaginaires depuis le début du XVIIe siècle. Dès 1629, A. Girard soupçonnait que toute équation de degré n a n racines réelles ou imaginaires, ce qui revenait à pressentir que les nombres imaginaires constituent le cadre « naturel » de la théorie des équations. À partir de 1675, Leibniz applique avec succès ses méthodes (de développements en série par exemple) aux nombres imaginaires et obtient ainsi de nombreux résultats, tandis qu'A. de Moivre, au début du XVIIIe siècle, met en évidence, par une utilisation systématique de la trigonométrie, les liens entre la recherche des racines des nombres imaginaires et la division d'un arc de circonférence en parties égales.

La possibilité, admise implicitement, d'étendre aux nombres complexes la plupart des notions relatives aux nombres réels allait se trouver mise en question par la *controverse des logarithmes* des nombres complexes, dont l'intérêt était apparu, par analogie avec le cas des pôles réels, dans l'intégration des fractions rationnelles. Les différentes formules contradictoires obtenues suscitèrent contre les imaginaires un vent de méfiance, qui fut dissipé par L. Euler ; celui-ci comprit qu'il fallait abandonner le caractère univoque du logarithme pour obtenir une théorie satisfaisante et établit d'innombrables formules relatives aux fonctions élémentaires d'une variable complexe.

À la fin du XVIIIe siècle, les imaginaires sont d'usage courant, mais leur « existence mathématique » véritable n'est pas établie ; c'est aux mathématiciens du XIXe siè-

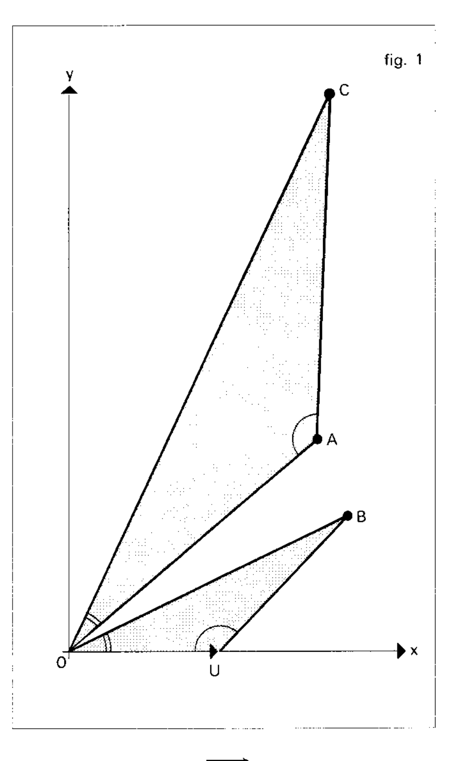
cle qu'il appartenait de les construire à partir des quantités connues, de leur donner une « réalité mathématique ». Avec Cauchy, c'est le prodigieux essor de la théorie des fonctions d'une variable complexe et le début de l'analyse contemporaine (cf. FONCTIONS ANALYTIQUES).

Théorie géométrique

L'idée, non seulement de représenter les nombres imaginaires par les points du plan, exprimée maladroitement par Wallis dès 1685, mais de les *définir* à partir de ces notions, est apparue dans deux mémoires, passés inaperçus à l'époque, du Danois Wessel (1798) et du Suisse Argand (1806). En fait, c'est Cauchy qui diffusera ce point de vue.

Dans le plan muni de deux axes de coordonnées Ox et Oy, on dira que les vecteurs d'origine O portés par Ox définissent les nombres réels, tandis que les autres vecteurs d'origine O définissent les nombres imaginaires ; le terme nombres complexes recouvre à la fois les nombres réels et les nombres imaginaires.

L'addition des nombres complexes se définit à partir de l'addition usuelle des vecteurs. Pour la multiplication, on fait la construction suivante : soit $\overrightarrow{OU}$ le vecteur unitaire de l'axe Ox ; le vecteur $\overrightarrow{OC}$ « produit » des vecteurs $\overrightarrow{OA}$ et $\overrightarrow{OB}$ s'obtient alors en construisant sur $\overrightarrow{OA}$ un triangle OAC directement semblable au triangle OUB (fig. 1). À partir de là, on retrouve géométriquement toutes les propriétés des nombres complexes.



Le nombre complexe i est défini par le vecteur unitaire de l'axe Oy, et la multiplication d'un vecteur $\overrightarrow{OA}$ par i revient donc à prendre un vecteur $\overrightarrow{OA'}$ directement perpendiculaire à $\overrightarrow{OA}$; répétant cette opération, on obtient le vecteur $\overrightarrow{OA''}$ opposé du vecteur $\overrightarrow{OA}$, c'est-à-dire que la multiplication par i^2 revient à multiplier par le nombre réel -1.

Théorie arithmétique

La théorie géométrique présente l'inconvénient de subordonner toutes les propriétés algébriques des nombres complexes à des considérations géométriques qui peuvent sembler étrangères. La théorie arithmétique, due à Hamilton (1835), consiste à considérer les nombres complexes comme des couples de nombres réels et à *définir* la somme et le produit par des formules explicites ; nous n'insisterons pas davantage ici sur cette approche, que nous exposerons ci-dessous. Ce point de vue conduit à essayer de définir plus généralement des opérations d'addition et de

multiplication pour des systèmes de n nombres réels et a conduit Hamilton à introduire les quaternions et plus généralement les algèbres de dimension finie (appelées, au XIXe siècle, *systèmes hypercomplexes* ; cf. ANNEAUX ET ALGÈBRES).

Les équivalences algébriques

Dès 1847, Cauchy considère que les calculs sur les nombres complexes reviennent à calculer sur les polynômes en la variable i, soumis aux règles usuelles de l'algèbre, en remplaçant $i^2 + 1$ par 0 ; cela revient à considérer que deux polynômes sont équivalents, c'est-à-dire définissent le même nombre complexe, si leur différence est divisible par $X^2 + 1$. Dans le langage contemporain, cela revient à définir l'ensemble des nombres complexes comme des classes d'équivalence de polynômes à coefficients réels modulo le polynôme irréductible $X^2 + 1$. C'est cette approche qui allait conduire Kronecker à la théorie générale des corps de nombres algébriques (cf. CORPS).

2. Le corps des nombres complexes

Construction

Par définition, un *nombre complexe* sera un couple $z = (x, y)$ de deux nombres réels ; si $z = (x, y)$ et $z' = (x', y')$ sont deux nombres complexes, on appelle alors somme et produit de ces deux nombres complexes les nombres complexes :

$$z + z' = (x + x', y + y')$$

et

$$zz' = (xx' - yy', xy' + x'y).$$

Il est alors facile de vérifier que, pour ces deux opérations, l'ensemble des couples de nombres réels est un corps, le *corps* **C** *des nombres complexes* ; par exemple, si $z = (x, y) \neq (0, 0)$, son inverse est le nombre complexe :

$$z^{-1} = \left(\frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2}\right)$$

aussi noté $1/z$.

Il est aussi souvent commode de représenter géométriquement le nombre complexe $z = (x, y)$ par le point M de coordonnées (x, y) dans le plan muni d'un système d'axes de coordonnées orthonormé ; M s'appelle l'*image* de z et il est clair que tout point M du plan est l'image d'un unique nombre complexe appelé *affixe* de M.

Les nombres complexes de la forme $(x, 0)$ forment un sous-corps de **C** qui est isomorphe au corps **R** des nombres réels par l'application qui à $(x, 0)$ fait correspondre x ; dans la suite, nous identifierons donc le nombre complexe $(x, 0)$ au nombre réel x, ce qui fait apparaître **R** comme un sous-corps de **C**. Parmi les nombres complexes, les nombres réels ont donc pour images les points de l'axe Ox, appelé pour cette raison *axe réel* ; remarquons que si a est un nombre réel, le produit de a et du nombre complexe $z = (x, y)$ est simplement $az = (ax, ay)$.

Le nombre complexe $i = (0,1)$ vérifie $i^2 = -1$, et tout nombre complexe $z = (x, y)$ peut s'écrire de manière unique $z = x + iy$, pour x et y réels ; nous adoptons désormais cette écriture, dite cartésienne, pour tout nombre complexe. Les nombres réels x et y s'appellent respectivement la *partie réelle* et la *partie imaginaire* de z et on note :

$$\mathrm{Re}\, z = x, \quad \mathrm{Im}\, z = y.$$

Les nombres complexes écrits sous forme cartésienne satisfont aux règles

usuelles du calcul élémentaire, en tenant compte du fait que $i^2 = -1$; par exemple :

$$(x + iy)(x' + iy') = xx' + ixy' + ix'y + i^2yy' = xx' - yy' + i(xy' + x'y),$$

ce qui redonne la formule ci-dessus du produit. Les nombres complexes non nuls de la forme $iy, y \in \mathbf{R}$, ont pour carré le nombre réel négatif $-y^2$; pour cette raison, ils sont dits *imaginaires purs* et l'axe Oy est appelé l'*axe imaginaire*.

On appelle *conjugué* du nombre complexe $z = x + iy$ le nombre complexe $\bar{z} = x - iy$; l'application de conjugaison, qui à tout nombre complexe fait correspondre son conjugué, est un automorphisme involutif du corps **C**, c'est-à-dire que l'on a :

$$\overline{z + z'} = \bar{z} + \bar{z}', \quad \overline{zz'} = \bar{z}\bar{z}', \quad \bar{\bar{z}} = z ;$$

il en résulte par exemple que si $z \neq 0$, alors $\bar{z} \neq 0$ et $\overline{1/z} = 1/\bar{z}$.

Pour tout nombre complexe z, le produit $N(z) = z\bar{z} = x^2 + y^2$ est un nombre réel positif ; c'est le carré de la distance de l'image de z à l'origine des coordonnées. On appelle *module* de z le nombre réel positif :

$$|z| = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}} ;$$

le module des nombres complexes possède les mêmes propriétés que la valeur absolue des nombres réels : $|z| = 0$ si et seulement si $z = 0$;

$$|z + z'| \leqslant |z| + |z'|; \quad |zz'| = |z||z'|.$$

Remarquons que l'inverse d'un nombre complexe non nul est égal à $1/z = \bar{z}/|z|^2$; en particulier, les nombres complexes de module 1, sur lesquels nous reviendrons, ont pour inverse leur conjugué.

Le théorème fondamental de l'algèbre

Les nombres complexes sont donc apparus très tôt comme le domaine naturel de la théorie des équations algébriques : toute équation algébrique peut être résolue dans ce corps. Plus précisément, le résultat fondamental est le suivant. Si P est un polynôme de degré n à coefficients complexes, il existe n nombres complexes $a_1, a_2, ..., a_n$, pas nécessairement distincts, tels que l'on ait identiquement :

$$P(z) = \lambda (z - a_1)(z - a_2) ... (z - a_n),$$

où λ est le coefficient du terme de plus haut degré. Ainsi, si l'on appelle ordre de multiplicité d'une racine le nombre de fois où elle apparaît dans la décomposition ci-dessus, tout polynôme de degré n a exactement n racines, chacune étant comptée avec son ordre de multiplicité.

Cette propriété était implicite pour de nombreux mathématiciens, mais c'est à d'Alembert que l'on doit la première tentative de démonstration, d'où le nom de théorème de d'Alembert que l'on donne souvent à cet énoncé. La démonstration de d'Alembert (1746) repose sur une argumentation analytique habile mais qui utilise des résultats de topologie. On doit à Euler (*Recherches sur les racines imaginaires des équations*, 1751) la première tentative de démonstration algébrique, qui fut reprise et améliorée tout au cours du XVIII[e] siècle par Lagrange, Laplace et d'autres. Mais ces démonstrations présentaient toutes des lacunes importantes. Gauss, dans sa dissertation de 1799, en fait l'historique critique et donne la première démonstration complète. Il reviendra à plusieurs reprises sur ce sujet et ne donnera pas moins de quatre démonstrations différentes du théorème fondamental de l'algèbre. Les développements de la théo-

rie des fonctions de variable complexe au XIXe siècle ont vu naître d'innombrables démonstrations de ce résultat (cf. FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe, chap. 4).

Un corps sur lequel tout polynôme se décompose en facteurs du premier degré est dit algébriquement clos ; cette propriété explique par exemple pourquoi la théorie des courbes algébriques se développe plus harmonieusement sur le corps des nombres complexes que sur le corps des nombres réels (cf. COURBES ALGÉBRIQUES).

Limites

Puisque le module des nombres complexes possède les mêmes propriétés que la valeur absolue des nombres réels, on peut définir de manière analogue toutes les notions relatives aux limites ; remarquons d'ailleurs que les définitions qui suivent, appliquées au cas particulier des nombres réels, redonnent toutes les notions correspondantes pour ces nombres.

On appelle suite de nombres complexes la donnée, pour tout entier naturel n, d'un nombre complexe z_n ; la suite correspondante est alors notée (z_n). On dit qu'une suite (z_n) de nombres complexes tend vers une limite u, ou converge vers u, pour n tendant vers l'infini, si pour tout nombre ε strictement positif, on a :

$$|z_n - u| < \varepsilon$$

pour n assez grand, c'est-à-dire sauf pour au plus un nombre fini d'entiers n. On écrit alors :

$$\lim_{n \to \infty} z_n = u.$$

La limite si elle existe est déterminée de manière unique, et toutes les propriétés usuelles des suites de nombres réels ne faisant pas intervenir la relation $\leqslant$ sont valables ici ; en particulier, le critère de Cauchy, qui permet d'affirmer qu'une suite est convergente sans connaître sa limite, s'applique.

Les *séries* de nombres complexes jouent un rôle absolument essentiel car elles interviennent dans la définition des fonctions analytiques d'une ou de plusieurs variables complexes, qui est une branche fondamentale de l'analyse. Soit (z_n) une suite de nombres complexes et soit (s_n) la suite des sommes partielles :

$$s_n = z_0 + z_1 + \dots + z_n ;$$

on dit que la série de terme général z_n, ou que la série :

$$\sum_n z_n,$$

est convergente de somme S, et on écrit :

$$\sum_n z_n = \mathrm{S},$$

si la suite (s_n) converge vers S. Les séries permettent de définir de nombreuses fonctions ; ainsi, la série de terme général $z^n/n\,!$ converge pour tout nombre complexe z et sa somme, la fonction exponentielle complexe :

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n\,!}$$

est sans conteste une des fonctions les plus importantes des mathématiques (cf. EXPONENTIELLE ET LOGARITHME). La règle de multiplication des séries permet d'établir la propriété fondamentale de la fonction exponentielle : si z et z' sont deux nombres complexes, on a :

$$(*) \qquad e^{z+z'} = e^z e^{z'}.$$

3. Forme trigonométrique

Trigonométrie

Les nombres complexes de module 1 peuvent être caractérisés comme les nombres complexes $\neq 0$ dont le conjugué et l'inverse sont égaux ; on vérifie facilement qu'ils forment un groupe multiplicatif que nous désignerons par U. Les images des éléments de U sont les points du cercle de centre O et de rayon 1 (appelé souvent « cercle trigonométrique ») ; l'application qui au nombre complexe $u \in$ U, d'image M, fait correspondre l'angle A(u) du demi-axe réel positif avec la demi-droite OM est un isomorphisme du groupe multiplicatif U sur le groupe additif des angles orientés de demi-droites et pourrait d'ailleurs servir à donner une définition rigoureuse de ces angles. L'étude du groupe U constitue ce qu'on appelle traditionnellement la trigonométrie ; l'outil pour définir de façon correcte les fonctions trigonométriques est la fonction exponentielle complexe.

Pour tout nombre réel t, le nombre complexe e^{it} appartient à U. En effet, on voit facilement sur le développement en série de e^{z} que le conjugué de e^{z} est $e^{\bar{z}}$ pour tout nombre complexe z ; on a donc, en utilisant aussi (*),

$$|e^{it}|^2 = e^{it}\,\overline{e^{it}} = e^{it}\,e^{-it} = e^0 = 1\,;$$

la formule (*) montre aussi que l'on a :

$$(**) \qquad e^{i(t+t')} = e^{it}\,e^{it'},$$

ce qui exprime que l'application qui au nombre réel t associe le nombre complexe $e^{it} \in$ U est un homomorphisme du groupe additif **R** sur le groupe multiplicatif U.

Par définition, on appelle cos t et sin t respectivement les parties réelle et imaginaire de e^{it}, soit :

$$e^{it} = \cos t + i \sin t\,;$$

puisque $|e^{it}| = 1$, on a $\cos^2 t + \sin^2 t = 1$ pour tout nombre réel t.

L'étude de e^{it} (cf. EXPONENTIELLE ET LOGARITHME) montre alors qu'il existe un nombre réel $\pi > 0$ tel que $e^{i\pi/2} = i$ et tel que l'application qui à t associe e^{it} soit une bijection de l'intervalle $[0, 2\pi[$ sur U. Puisque, d'après (*) :

$$e^{2i\pi} = (e^{i\pi/2})^4 = i^4 = 1,$$

on en déduit, toujours d'après (*), que la fonction e^{it} est périodique de période 2π. Ainsi, tout nombre complexe u de module 1 s'écrit sous la forme :

$$u = e^{it} = \cos t + i \sin t,$$

où t est un nombre réel déterminé à $2k\pi$ près, k entier relatif ; cela revient à dire que si x et y sont deux nombres réels tels que $x^2 + y^2 = 1$, il existe un nombre réel t, défini à $2k\pi$ près, tel que $x = \cos t$ et $y = \sin t$. La propriété (**) montre, d'autre part, que si t et t' sont deux nombres réels, on a :

$$(\cos t + i \sin t)(\cos t' + i \sin t') = \cos(t + t') + i \sin(t + t'),$$

ce qui, en égalant les parties réelles et imaginaires des deux membres, donne les formules trigonométriques d'addition des arguments. On déduit facilement de ce qui précède la formule de De Moivre, valable pour tout entier relatif n,

$$(\cos t + i \sin t)^n = \cos nt + i \sin nt,$$

qui permet d'obtenir de nombreuses formules de trigonométrie.

Forme trigonométrique

Nous désignerons par **C*** le groupe multiplicatif des nombres complexes non nuls. Si $z \neq 0$, le nombre complexe $z/|z|$ est de

module 1 et on voit facilement que l'application qui à tout nombre complexe $z \neq 0$ associe le couple $(|z|, z/|z|)$ est une bijection de $\mathbf{C}^*$ sur l'ensemble $\mathbf{R}_+^* \times \mathrm{U}$ des couples (r, u) d'un nombre réel $r > 0$ et d'un élément $u \in \mathrm{U}$; la bijection réciproque associe à un tel couple (r, u) le nombre complexe ru, de module r. L'étude de U faite ci-dessus permet donc d'écrire tout nombre complexe $z \neq 0$ sous la forme :

$$z = re^{it} = r(\cos t + i \sin t), \text{ où } r = |z|,$$

appelée *forme trigonométrique* du nombre complexe z. On dit qu'une telle valeur de t est un *argument* de z ; quand on connaît une valeur de l'argument, on obtient donc toutes les autres en lui ajoutant un multiple entier relatif de 2π et deux nombres complexes sont égaux si et seulement s'ils ont le même module et des arguments qui diffèrent de $2k\pi$, k entier relatif. Si on impose à l'argument d'appartenir à l'intervalle $]-\pi, +\pi]$, il est déterminé de manière unique et s'appelle l'*argument principal.*

Racines n-ièmes

La recherche des nombres complexes z tels que $z^n = 1$ va montrer l'intérêt de la forme trigonométrique. Écrivant z sous forme trigonométrique :

$$z = r(\cos t + i \sin t),$$

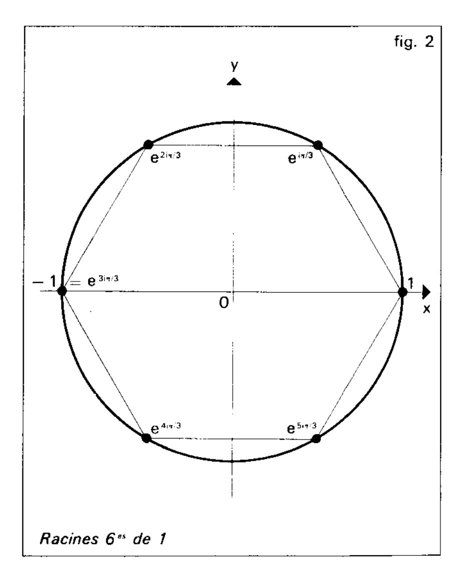
on doit avoir :

$$z^n = r^n(\cos t + i \sin t)^n = r^n(\cos nt + i \sin nt) = 1 ;$$

les nombres z^n et 1 sont égaux s'ils ont le même module, soit $r^n = 1$, d'où $r = 1$, et s'ils ont des arguments qui diffèrent d'un multiple entier de 2π, soit $nt = 2k\pi$, avec k entier relatif, en prenant 0 pour argument de 1. On peut donc écrire $t = 2k\pi/n$; si on se limite aux valeurs de t appartenant à l'intervalle $[0, 2\pi]$, il suffit de donner à k les n valeurs successives 0, 1, 2, ..., $n-1$, car toute autre valeur de k donne des valeurs correspondantes de t égales, à un multiple de 2π près, aux valeurs de t obtenues pour ces nombres. On obtient ainsi n nombres complexes de module 1 distincts :

$$z_k = \cos \frac{2k\pi}{n} + i \sin \frac{2k\pi}{n},$$

où $k = 0, 1, 2, ..., n-1$; ces nombres complexes sont les n racines du polynôme $z^n - 1$ et leurs images sont les sommets d'un polygone régulier inscrit dans le cercle trigonométrique (fig. 2).



Racines 6es de 1

Un raisonnement analogue montrerait que tout nombre complexe non nul a n racines n-ièmes distinctes ; si c est l'une d'entre elles, ces racines sont :

$$cz_0, cz_1, ..., cz_{n-1},$$

où les z_k sont les racines n-ièmes de 1.

JEAN-LUC VERLEY

Bibliographie

R. Argand, *Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques*, nouv. tir., Blanchard, Paris, 1971 / G. Chevalier, J.-F. Fournier & J. Siméon, *Les Nombres complexes*, Université scientifique, technologique et médicale de Grenoble, Saint-Martin-d'Hères, 1987 / J. Trignan, *La Géométrie des nombres complexes*, Bréal, 1991.

CONIQUES

L'étude des coniques a été pendant deux millénaires le terrain de prédilection des géomètres qui ont accumulé sur ce sujet d'innombrables théorèmes. Dès la fin du IIIe siècle avant J.-C., les mathématiciens avaient obtenu par des méthodes purement géométriques des résultats très complets : le *Traité des sections coniques* d'Apollonius (né vers 245 avant J.-C.) est un des sommets de la mathématique grecque.

Le XVIIe siècle allait voir à nouveau se développer la théorie des coniques dans deux directions très différentes. Descartes met en évidence les équations des coniques et reconnaît qu'elles constituent les courbes du second degré, tandis que Pascal et Desargues donnent une impulsion considérable à la géométrie pure en inaugurant l'étude projective des coniques.

De nos jours, les mathématiciens ne se préoccupent plus guère d'enrichir l'herbier un peu vieillot des théorèmes sur les coniques, qui ont été réduites à un chapitre de la théorie des formes quadratiques. Une conique apparaît aujourd'hui comme une courbe non vide du plan projectif, définie par une équation $Q(x, y, t) = 0$, où Q est une forme quadratique en les coordonnées homogènes x, y, t ; cette définition est la seule qui contienne tous les cas particuliers et elle s'étend immédiatement en dimension supérieure aux quadriques et aux hyperquadriques.

On se limite dans ce qui suit à des résultats purement classiques, en renvoyant à l'article formes QUADRATIQUES pour un exposé plus moderne.

E. U.



1. Les sections coniques

Le cône circulaire

Le cercle est la figure conique la plus simple et la plus ancienne ; il a été considéré comme une figure bien avant le couple de droites, pourtant plus simple a priori (de tels couples existent dans toute géométrie, alors que le cercle n'apparaît que dans quelques-unes). On peut alors définir le cône circulaire, ensemble des droites s'appuyant sur un point fixe (le sommet O) et sur un cercle (la base C). Le plus simple de ces cônes est le cône de révolution, où la droite qui joint O au centre de C est perpendiculaire au plan du cercle.

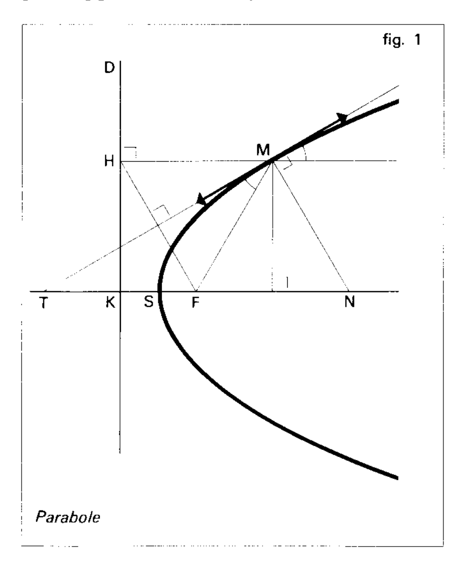
Les sections d'un cône circulaire par un plan sont appelées sections coniques. On peut ainsi obtenir, outre le cercle, des ellipses, des paraboles, des hyperboles et des figures particulières (droites sécantes si le plan passe par le sommet, droites confondues s'il contient de plus une tangente au cercle). Seul échappe à cette définition (conforme à l'étymologie) le cas de deux droites parallèles distinctes. Celui-ci pourra néanmoins venir compléter la famille en application du théorème : *Toute section plane d'un cône dont une base est une conique est elle-même une conique ou le plan tout entier.*

Ce théorème capital, qui va beaucoup plus loin que la définition grecque (qui ne considérait que certains types de cônes), affirme en quelque sorte que la notion de conique est la notion projective fondamentale, c'est-à-dire la notion invariante dans [illegible] considérations affines et surtout métriques

que les coniques ont été reconnues et étudiées. Dans un plan métrique tel que le plan euclidien traditionnel, les cas particuliers de coniques décomposées sont triviaux et on ne les considérera plus dans le cadre de cet article. Il suffira de [illegible] cercle).

Étant donné une droite D, appelée directrice, et un point F (le foyer) non situé sur elle, la parabole est :
- l'ensemble des points M qui sont centre d'un cercle passant par F et tangent à la droite D (définition 1) ;
- l'ensemble des points M tels que la distance MF soit égale à la distance MH de M à la droite D (définition 2).

La perpendiculaire à D issue de F est l'axe de la parabole. Si elle coupe la directrice en un point K, la distance $FK = p$ est le paramètre de la parabole. Le sommet S, situé sur la courbe, est le milieu de KF ; la médiatrice de KF est d'ailleurs la tangente au sommet. Toute la courbe est connexe, convexe et symétrique par rapport à l'axe (fig. 1).



Parabole

Tangentes

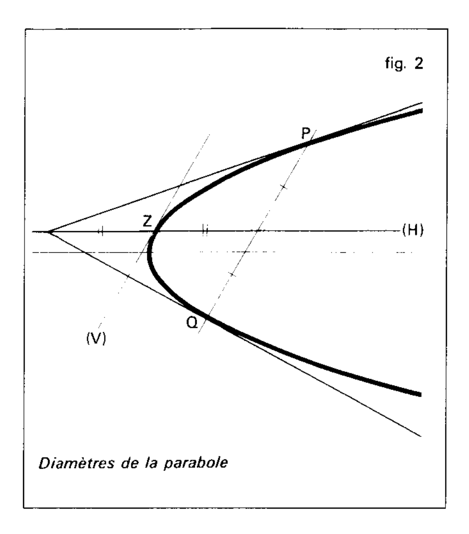
En chaque point M de la parabole, il existe une tangente. Celle-ci est bissectrice de l'angle formé par MF et la parallèle à l'axe ; cette bissectrice rencontre l'axe en un point T tel que S soit milieu de la projection sur l'axe du segment MT : cela détermine entièrement la tangente en M (fig. 1). La normale coupe l'axe en un point N tel que les vecteurs $\overrightarrow{KF}$ et $\overrightarrow{MN}$ aient des projections de même valeur sur l'axe : l'invariance de la projection de MN est une propriété caractéristique de la parabole.

Tout point d'une parabole pouvant être pris comme sommet si l'on modifie convenablement la condition d'orthogonalité dans le plan, certaines des propriétés précédentes sont en fait affines et non réellement métriques. Prenons par exemple une parallèle quelconque (H) à l'axe. Elle coupe la parabole en un point unique Z où la tangente sera notée (V) : on constate alors que la parabole est invariante dans la symétrie par rapport à (H) parallèlement à (V), et que Z est le milieu de la projection de la partie de tangente à la parabole comprise entre son point de contact et (H) [fig. 2]. (H) est appelé diamètre de la direction de (V), c'est-à-dire ensemble des milieux des segments PQ parallèles à (V) dont les extrémités sont sur la parabole ; les tangentes en P et Q se coupent d'ailleurs sur (H). La démonstration de ces propositions peut se ramener à une simple projection d'une parabole sur un autre plan, Z étant alors l'image du sommet, (H) celle de l'axe, etc.

Le dénombrement des points d'intersection d'une droite avec une parabole, la partition du plan en « extérieur » et « intérieur » sont naturellement des notions affines. On se bornera pourtant, pour la commodité, à considérer une parabole métrique, la projection signalée ci-dessus permettant l'extension de ces notions à une parabole affine.

Intersection avec une droite

Une droite coupe une parabole en un ou deux points. Si le point d'intersection est



Diamètres de la parabole

unique et la droite non parallèle à l'axe, la droite est tangente. Pour qu'il en soit ainsi, il faut et il suffit que la projection du foyer F sur elle appartienne à la tangente au sommet, ce qui donne une définition traditionnelle de la parabole comme enveloppe de droites (dont la projection sur elles d'un point fixe décrit une droite fixe). Le point de contact est alors aisément déterminé par sa projection sur l'axe, symétrique de T par rapport à S. Il existe une tangente unique ayant une direction donnée, sauf si celle-ci est celle de l'axe. Si une tangente variable coupe deux tangentes fixes distinctes en U et U′, les abscisses de ces points sont liées par une relation affine ($x' = ax + b$), ce qui est une propriété affine, et le triangle UFU′ reste semblable à lui-même (ses angles sont constants), ce qui est une propriété métrique ; toutes deux sont caractéristiques de la parabole. Si les tangentes fixes se coupent en V, F appartient au cercle (UVU′) et l'orthocentre du triangle UVU′ est sur la directrice.

Donnons-nous une droite quelconque D. Si la projection de F sur D est sur la tangente au sommet, on vient de voir que D est une tangente à la parabole. Si cette projection est du même côté de cette tangente au sommet que F, D coupe effectivement la parabole en deux points distincts (un seul si D est parallèle à l'axe) que l'on peut construire à la règle et au compas (suivant les vieilles règles grecques). Si cette projection est située dans l'autre demi-plan, il n'existe aucun point d'intersection.

Prenant le problème d'un point de vue différent, on peut chercher les tangentes à la parabole issues d'un point donné M. Il en existe une seule si M est sur la courbe. Sinon MF > MH définit l'extérieur de la parabole, d'où l'on peut mener deux tangentes à la courbe (perpendiculaires si M appartient à la directrice) ; de l'intérieur, caractérisé par MF < MH, on ne peut mener de tangentes. Le foyer est à l'intérieur.

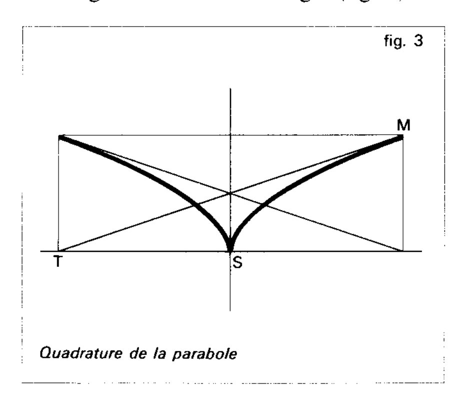
Propriétés différentielles et intégrales

Par un point donné, il peut exister jusqu'à trois normales à la parabole, coupant celle-ci en trois points définissant un cercle passant par le sommet, et un triangle dont le centre de gravité appartient à l'axe.

Le centre de courbure C en M, point où la normale en M est tangente à son enveloppe, se projette en P sur MF de façon que FM = FP = FN (N étant l'intersection de la normale et de l'axe) ; CM = R est le rayon de courbure, lié à MF par l'égalité $pR^2 = 8MF^3$. La normale CNM coupe la directrice en D tel que $\overrightarrow{CM} = 2\,\overrightarrow{MD}$.

L'aire comprise entre un arc de parabole SM, l'axe et la projetante de M sur

l'axe est égale aux deux tiers de l'aire du rectangle de diagonale SM, de côtés parallèles et perpendiculaires à l'axe (une telle propriété se conserve, *mutatis mutandis*, par projection). Cette quadrature, la première du genre, était déjà connue d'Archimède qui utilisait là sa méthode d'exhaustion (cf. CALCUL INFINITÉSIMAL). Des géomètres tels que Roberval retrouvèrent ce résultat par des raisonnements inspirés de la mécanique et du fait que la portion de tangente comprise entre M et l'axe était coupée en son milieu par sa tangente au sommet : ils obtenaient ainsi deux arcs symétriques de paraboles découpant trois aires égales dans un rectangle (fig. 3).



Quadrature de la parabole

La plupart des propriétés de la parabole sont évidemment généralisables aux autres coniques, dites coniques à centre ; pour certaines d'entre elles, la parabole apparaît en quelque sorte comme un moyen terme entre l'ellipse et l'hyperbole. Mais la parabole garde son originalité : par exemple, ce qui est rare pour une conique autre qu'un cercle, la parabole est rectifiable, ce qui veut dire que l'on peut donner une formule explicite pour la longueur d'un arc de parabole SM. Dans un système d'axes orthogonaux d'origine S dont l'axe des abscisses est l'axe de la parabole, si t est un nombre tel qu'abscisse et ordonnée de M soient égales respectivement à :

$$x = (p/2)\,\mathrm{sh}^2 t,\ y = p\,\mathrm{sh}\,t \quad (\text{d'où } y^2 = 2px),$$

alors la longueur de l'arc SM est :

$$s = (p/4)\,(2t + \mathrm{sh}\,2t).$$

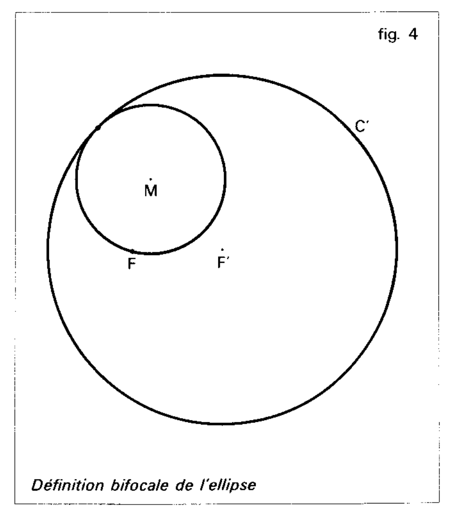
Les longueurs des arcs des autres coniques ne peuvent s'exprimer en général qu'à l'aide de fonctions elliptiques.

3. Les coniques à centre

Foyers

Pour les coniques à centre, les deux définitions métriques courantes sont les suivantes :

– l'ensemble des points M qui sont centre d'un cercle passant par un point donné F et tangent à un cercle donné C′ de centre F′ : définition bifocale (fig. 4) ;



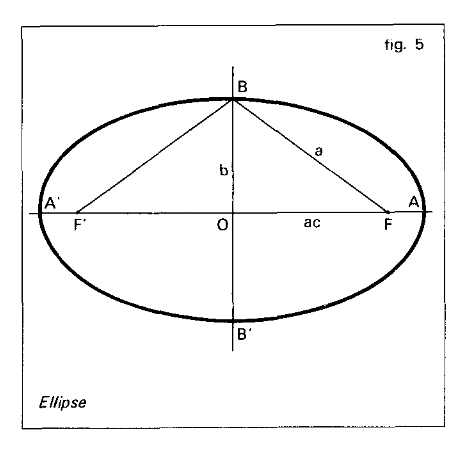
Définition bifocale de l'ellipse

– l'ensemble des points M tels que la distance MF soit proportionnelle à la distance MH de M à la droite D (directrice associée à F) : définition monofocale.

La première met en jeu deux foyers, F et F′, et le cercle directeur C′ de centre F′ et de rayon ($2a$). Par hypothèse, FF′ $= 2c$ est distinct de $2a$ (F n'appartient pas à C′). Si l'on peut écrire $c < a$, alors F est intérieur à C′ et la courbe obtenue, lieu de M tel que :

$$MF + MF' = 2a,$$

est une ellipse, convexe, connexe et bornée (fig. 5). Si $c > a$, F est extérieur à C′ :



Ellipse

l'hyperbole obtenue est composée de deux parties convexes et connexes respectivement définies par $MF - MF' = 2a$ et $MF' - MF = 2a$. Pour un cercle, $c = 0$.

Ces coniques admettent deux axes de symétrie perpendiculaires (dont FF′, appelé axe focal) et un centre de symétrie O, milieu de FF′. L'axe focal coupe la courbe aux sommets A et A′, tels que :

$$OA = OA' = a.$$

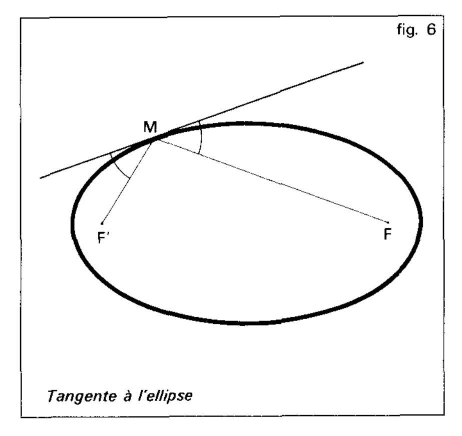
Une ellipse a deux autres sommets B et B′ sur l'axe secondaire de symétrie, définis par $OB = OB' = b$ avec

$$a^2 - c^2 = b^2 \text{ (fig. 5).}$$

Pour une hyperbole, on pose au contraire : $c^2 - a^2 = b^2$, si l'on veut n'utiliser que des nombres réels.

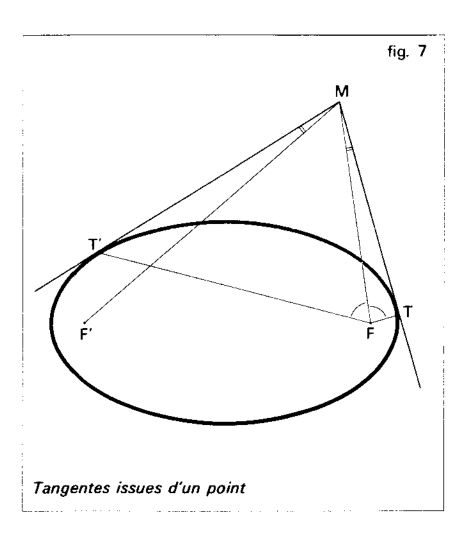
Tangentes

En chaque point d'une conique à centre il existe une tangente. Celle-ci est la bissectrice de l'angle géométrique $\widehat{FMF'}$ (extérieure pour l'ellipse, fig. 6, intérieure pour



Tangente à l'ellipse

l'hyperbole). On ne peut mener deux tangentes MT et MT′ distinctes à la conique que par un point M extérieur à celle-ci, défini par exemple par la relation $MF > eMH$, où $e = c/a$ est l'excentricité de la conique (inférieure à 1 pour l'ellipse, égale à 1 pour la parabole par définition, et supérieure à 1 pour l'hyperbole). Sur la conique même ($MF = eMH$), ce qui est la seconde définition, il y a une tangente unique (double). D'un point intérieur ($MF < eMH$) on ne peut pas mener de tangentes ; c'est notamment le cas en F ou en F′. Si les tangentes existent, MT et MT′ ont mêmes bissectrices que MF et MF′ ; de plus FM est bissectrice de l'angle $\widehat{TFT'}$ (cf. fig. 7, pour l'ellipse). MT est perpendiculaire à MT′ si M appartient au cercle de Monge, de centre O et de rayon $\sqrt{2a^2 - c^2}$ (pour $e \leqslant \sqrt{2}$).



Tangentes issues d'un point

La projection de F sur une tangente décrit le cercle de diamètre AA' ; si la projection de F sur une droite est dans la même région que F par rapport à ce cercle, la droite donnée coupe la conique en deux points distincts que l'on peut construire ; si la projection est dans l'autre région, la droite ne coupe pas la conique. On peut déduire de cela des définitions de l'ellipse et de l'hyperbole comme enveloppes de droites, comme pour la parabole.

Parmi les tangentes à une hyperbole, deux sont particulières. Issues de O, elles n'ont aucun point de contact à distance finie avec la courbe dont elles sont des asymptotes ; elles déterminent deux angles contenant chacun une branche connexe de l'hyperbole.

Correspondances linéaires

Si une tangente variable coupe deux tangentes fixes à une conique à centre, les abscisses x et x' des deux points d'intersection sont reliées par une relation homographique du type :

$$axx' + bx + cx' + d = 0.$$

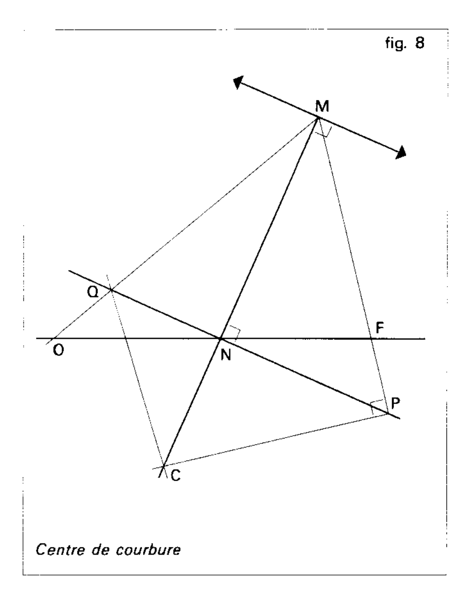
Il existe une réciproque importante à cette propriété affine : une droite joignant deux points de deux droites fixes dont les abscisses sont reliées par une telle relation (où $a \neq 0$) enveloppe une conique à centre ($a = 0$ correspond à une parabole). Le produit des distances des deux foyers à une tangente variable est égale à b^2 ; les foyers sont de part et d'autre pour une hyperbole, du même côté pour une ellipse. Certaines des propriétés précédentes s'étendent naturellement à une parabole, en supposant par exemple que la direction de F' est celle de l'axe.

Pour que quatre points d'une conique soient sur un même cercle, il faut et il suffit que deux des cordes les joignant soient symétriques par rapport aux axes de la conique. D'un point du plan, on peut mener jusqu'à quatre normales à une conique à centre. Les pieds de trois d'entre elles et le point diamétralement opposé au quatrième sont sur un même cercle.

Si l'on se donne deux directions perpendiculaires, les points d'où les tangentes issues ont des directions symétriques par rapport à ces directions privilégiées appartiennent à une hyperbole (équilatère) de même centre que la conique, passant par les deux foyers ; il en serait de même si l'on remplaçait les deux directions choisies par celles de leurs bissectrices, d'où la position des foyers à l'intersection des deux hyperboles.

Propriétés différentielles et intégrales

La normale en M coupe l'axe focal en N tel que NF $= e$MF. La perpendiculaire en N à MN coupe MF en P, projection du centre de courbure C en M sur MF, et coupe MO en Q, point tel que QC soit perpendiculaire à l'axe focal (fig. 8) : cela



Centre de courbure

fournit deux constructions du centre de courbure, sauf aux sommets de l'axe focal où le rayon de courbure R = MC est égal à $p = b^2/a$, paramètre de la conique (et longueur de la demi-corde issue de F et perpendiculaire à FF′). En un point quelconque, on a :

$$a^2b^2R^2 = MF^3MF'^3, \quad p^2R = MN^3.$$

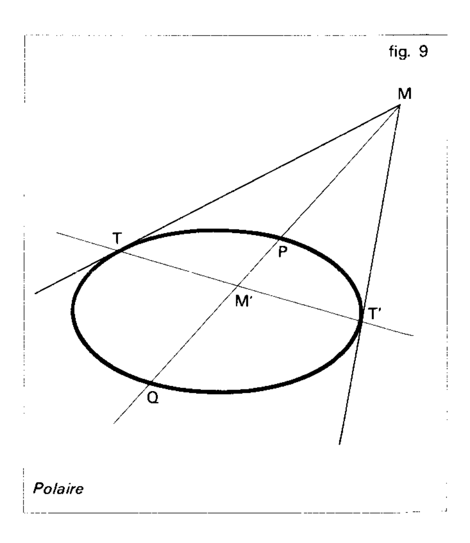
On retrouve aussi le paramètre comme longueur de la projection orthogonale de MN sur MF ou MF′. Si N′ est sur la médiatrice de FF′ et sur MN, MN′ se projette suivant une longueur égale à a sur MF ou MF′, et les trois vecteurs $\overrightarrow{N'N}$, $\overrightarrow{N'M}$ et $\overrightarrow{NM}$ sont proportionnels aux nombres c^2, a^2 et $a^2 - c^2$ (b^2 pour une ellipse, $-b^2$ pour une hyperbole).

Les centres de courbure en les sommets A et B d'une ellipse sont alignés avec le point d'intersection des tangentes en A et B ; cette droite est perpendiculaire à AB ; la construction des cercles de courbure en A et B permet une construction graphique très soignée de l'arc AB.

Pôles et polaires

Si une corde MM′ coupe la directrice D en P, FP est une bissectrice de l'angle $\widehat{MFM'}$. Si MM′ passe par F, les tangentes en M et en M′ se coupent en T sur la directrice, et P est conjugé harmonique de F par rapport à M et M′ ; de plus l'angle $\widehat{MFT}$ est droit. Cette directrice D est située à une distance a^2/c de O, est perpendiculaire à l'axe focal et située du même côté de O que F. Elle n'existe pas pour un cercle pour lequel F, F′ et O sont confondus, d'où $c = e = 0$; cela fait que la définition monofocale, qui masque déjà le caractère de double symétrie de la conique, n'est pas aussi générale qu'elle paraît. Il existe naturellement une directrice analogue D′ pour l'autre foyer F′.

Pour une conique propre quelconque, à centre ou parabole, les conjugués harmoniques d'un point fixe M par rapport aux extrémités P et Q d'une corde passant par M sont situés sur une droite, appelée polaire de M (fig. 9). Ainsi une directrice est la polaire du foyer correspondant. Seul



Polaire

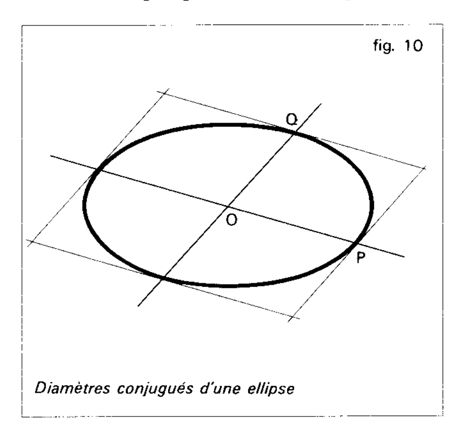
le centre n'a pas de polaire. Un point de la conique a sa tangente comme polaire. Toute droite ne passant pas par le centre éventuel est la polaire d'un certain point qui est appelé son pôle.

Si la polaire de M passe par M′, celle de M′ passe par M. Si M, P, Q sont alignés, leurs polaires sont concourantes et réciproquement. Si MT et MT′ sont des tangentes issues de M, la corde TT′ qui joint leurs points de contact est la polaire de M (cf. fig. 9 pour l'ellipse). Ces propriétés peuvent être obtenues très simplement par perspective (conique éventuellement) à partir des propriétés analogues pour le cercle, où elles sont bien connues ; elles sont projectives en dépit des apparences.

Diamètres

Donnons-nous une direction D ; les milieux des cordes parallèles à D sont situés sur un diamètre, c'est-à-dire une droite passant par le centre (ellipse ou hyperbole) ou parallèle à l'axe (parabole). La conique est invariante dans la symétrie par rapport à ce diamètre parallèlement à la direction D. Fixons un point M ; à tout diamètre de direction D, associons son point d'intersection avec la perpendiculaire issue de M au diamètre conjugué de D défini ci-dessus. Ces points engendrent une hyperbole (équilatère), coupant la conique aux points où la normale passe par M, ce qui permet de les construire.

Pour une ellipse, à tout diamètre D, on peut associer le diamètre conjugué D′ dont D est à nouveau le conjugué. D et D′ coupent l'ellipse en P et Q, par exemple, tels que la tangente en P soit parallèle à D′, et forme avec D, D′ et la tangente en Q, un parallélogramme d'aire constante, égale à ab, comme on le voit en prenant les axes de l'ellipse pour D et D′ (fig. 10). De



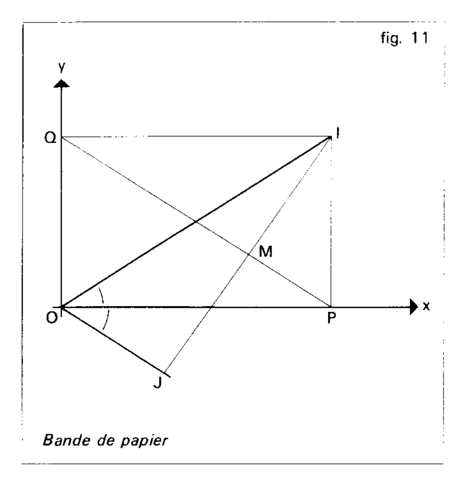
Diamètres conjugués d'une ellipse

plus $OP^2 + OQ^2 = a^2 + b^2$ est constant. De tels diamètres conjugués sont simplement les projections orthogonales de deux diamètres perpendiculaires d'un cercle de diamètre AA′ placé de façon que l'ellipse soit sa projection : la plupart des propriétés de la conjugaison (ainsi que la formule donnant l'aire de l'ellipse : $S = \pi ab$) sont des conséquences immédiates de cette projection.

4. Propriétés particulières

L'ellipse

Une construction très simple permet d'obtenir autant de points de l'ellipse que l'on en désire. Donnons-nous deux droites perpendiculaires Ox et Oy, et un segment constant PQ dont les extrémités décrivent respectivement Ox et Oy (fig. 11). Un point M situé entre P et Q tel que $MQ = a$ et $MP = b$ décrit une ellipse de centre O, d'axe focal Ox. La tangente en M est perpendiculaire à IM, où I est le quatrième sommet d'un rectangle OPIQ. Deux positions perpendiculaires de PQ (appelé



« bande de papier » dans la littérature) correspondent à deux extrémités de diamètres conjugués. I décrit le cercle de centre O et de rayon $(a + b)$; le symétrique J de I par rapport à M est lui aussi sur la normale en M ; il décrit le cercle de centre O et de rayon $(a - b)$, et Ox est bissectrice intérieure de l'angle $\widehat{IOJ}$; IJFF' sont sur un même cercle, et forment la figure connue sous le nom de quadrangle harmonique, inverse d'une division harmonique (fig. 11).

La projection considérée, que l'on peut relier à l'affinité d'axe AA' et de rapport b/a, qui transforme le cercle de diamètre AA' en l'ellipse, est à l'origine d'une représentation paramétrique particulièrement simple de celle-ci. Le transformé du point du cercle définissant, avec Ox, un angle t est en effet le point de l'ellipse de coordonnées $x = a \cos t$, $y = b \sin t$; t est l'anomalie excentrique de ce point. Deux extrémités de diamètres conjugués, dont les pentes dans ce système d'axes ont pour produit b^2/a^2, ont des anomalies excentriques différentes d'un angle droit. L'équation de l'ellipse en découle ; on pourrait aussi la déduire des formules d'Euler, valables également pour une hyperbole :

$$MF = |a - ex|, \ MF' = |a + ex|$$

pourvu que F soit le point de coordonnées $(c, 0)$. Cette équation s'écrit :

$$x^2/a^2 + y^2/b^2 = 1.$$

L'ellipse intervient notamment en astronomie. Tout point attiré par un autre point F suivant la loi de Newton décrit une conique de foyer F. Les planètes, qui ont un mouvement périodique, décrivent des ellipses, seules coniques à être bornées. L'aire comprise entre la courbe et les droites joignant F à deux positions du point mobile est proportionnelle à l'intervalle de temps séparant les deux positions ; pour une planète, la période T du mouvement est telle que T^2 soit proportionnel au cube a^3 de la longueur :

$$a = OA = AA'/2\,;$$

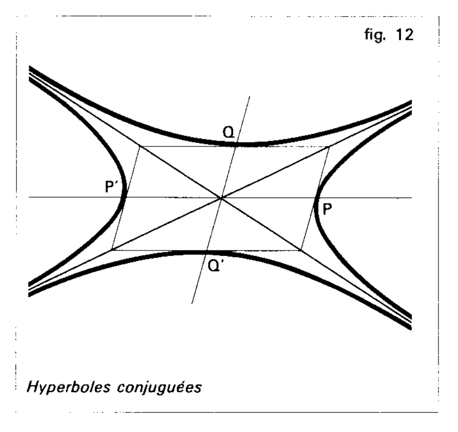
ce sont les fameuses lois de Kepler.

L'hyperbole

L'hyperbole, à cause de ses asymptotes, possède des propriétés très particulières. Il n'existe pas nécessairement de tangentes ayant une direction donnée (il suffit de considérer des droites passant par le centre et situées dans les deux angles déterminés par les asymptotes où sont situées les deux branches). L'angle de ces asymptotes est l'angle $2z$ défini par $\cos z = 1/e$. Les points d'intersection avec les tangentes en A et A' sont les sommets d'un rectangle de côtés $2a$ et $2b$.

Cette dernière propriété se généralise de la façon suivante. Considérons un diamètre P'OP de l'hyperbole. La droite passant par O et parallèle à la tangente en P est le diamètre conjugué de la direction OP. Il ne coupe pas l'hyperbole. Mais si

l'on y construit le point Q tel que $OP^2 - OQ^2 = a^2 - b^2$, l'aire du triangle OPQ est constante (et vaut $ab/2$) ; quand P varie, Q décrit une autre hyperbole, dite conjuguée de la précédente, ayant même centre et mêmes asymptotes, échangeant a et b avec la précédente (fig. 12). De plus, la tangente



Hyperboles conjuguées

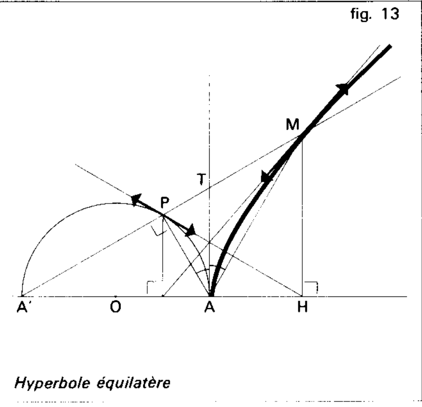
en Q est parallèle à OP, et coupe la tangente en P sur une asymptote. Enfin la conjuguée de la conjuguée est l'hyperbole originale. Les deux diamètres « conjugués » OP et OQ forment un faisceau harmonique avec les asymptotes. La symétrie par rapport à OP et parallèlement à OQ laisse invariante chacune des deux hyperboles conjuguées.

Revenant à une hyperbole simple, la famille de ses tangentes jouit d'une propriété remarquable (due au fait que les asymptotes sont des tangentes très particulières) ; partant du fait qu'une corde MM' coupe les asymptotes en des points P et P' tels que l'on ait $\overrightarrow{MP} = \overrightarrow{P'M'}$, la tangente en M coupe les asymptotes en Q et Q' tels que M soit le milieu de QQ'. QQ'FF' sont situés sur un même cercle et forment un quadrangle harmonique. De plus, si une droite passant par M coupe les asymptotes en P et P' et se déplace parallèlement à elle-même, le produit MP . MP' reste constant : il est notamment égal à b^2 si MPP' est perpendiculaire à l'axe focal.

L'hyperbole équilatère

Parmi toutes les hyperboles, l'hyperbole équilatère, d'excentricité $\sqrt{2}$ ($a = b$), est particulièrement intéressante. Ses asymptotes sont perpendiculaires.

Considérons les sommets A et A', le cercle de diamètre AA', un point P décrivant ce cercle (fig. 13). Menons par A la



Hyperbole équilatère

droite symétrique de AP par rapport à la direction de AA' ; cette droite coupe A'P en M, point de la branche contenant A de l'hyperbole équilatère de sommets A et A'. Le triangle AMA' est pseudo-rectangle, c'est-à-dire que la différence des angles en A et A' est égale à un droit. La tangente en M au cercle (AMA') est perpendiculaire à AA', qu'elle coupe en H tel que $HM^2 = HA . HA'$ (comme dans un véritable triangle rectangle). Si A'PM coupe la tangente au sommet A en le point T, conjugué harmonique de A' par rapport à M et P, les tangentes en P (au cercle) et en M (à l'hyperbole) se coupent sur AT au milieu de ce segment ; elles coupent AA' en les projections de P et de M ; M se projette

sur AT en un point appartenant à OP, etc. (fig. 13). Cette figure est la base de la correspondance entre les trigonométries circulaire et hyperbolique ; elle inspira à Abraham de Moivre sa fameuse formule :

$$(\cos x + i \sin x)^n = \cos nx + i \sin nx.$$

Si l'aire comprise entre OA, OM et l'hyperbole est égale à $a^2t/2$, les coordonnées de M (dans les axes convenables) sont alors $x = a$ ch t, $y = a$ sh t, d'où l'équation $x^2 - y^2 = a^2$. Toute hyperbole étant affine d'une hyperbole équilatère de même sommet, l'équation générale d'une hyperbole est donc :

$$x^2/a^2 - y^2/b^2 = 1,$$

et celle des asymptotes est obtenue en y remplaçant 1 par 0. La fonction « gudermanien de t », utilisée pour construire les cartes géographiques selon la projection de Mercator, est définie par :

$$t' = 2 \operatorname{arc\,tg} e^t - \pi/2.$$

$a^2/2.t'$ est la moitié de l'aire comprise entre OA, OP et le cercle.

Dans une hyperbole équilatère :

$$MO^2 = MF \,.\, MF' = MN^2,$$

le rayon de courbure R est tel que $a^2R = OM^3$. Si C et C′ sont deux points de l'hyperbole symétriques par rapport à O, les bissectrices de $\widehat{CMC'}$ sont parallèles aux asymptotes, et la différence des angles C et C′ est constante. Tout cercle passant par CC′ est recoupé suivant un diamètre ; un cercle de centre C et de rayon CC′ recoupe l'hyperbole équilatère suivant les sommets d'un triangle équilatéral. Deux hyperboles équilatères se coupant en les sommets et l'orthocentre d'un triangle, on peut en déduire de nombreuses propriétés. Si une corde MM′ se déplace en restant parallèle à elle-même, en coupant toujours l'hyperbole, le cercle de diamètre MM′ passe par deux points fixes où la normale est parallèle à MM′.

ANDRÉ WARUSFEL

Bibliographie

APOLLONIUS DE PERGA, *Les Coniques* (Κωνικά), P. Vereecke éd., Paris, 1963 / M. BERGER, *Géométrie*, t. IV : *Formes quadratiques, quadriques et coniques*, Cedic-F. Nathan, Paris, 1978 / J. HADAMARD, *Leçons de géométrie élémentaire*, 2 vol., A. Colin, 1949, repr. J. Gabay, Sceaux, 1988 / H. LEBESGUE, *Les Coniques*, J. Gabay, Sceaux, 1988 / J. PICHON, *Géométrie analytique, coniques*, Ellipses, 1990.

CONVEXITÉ

La convexité, étude des ensembles et des fonctions convexes, constitue une branche de la géométrie et de l'analyse qui unifie des phénomènes à première vue totalement dissemblables. Elle intervient à divers niveaux dans des branches très variées des mathématiques : théorie des nombres, problèmes combinatoires, analyse fonctionnelle et applications (théorie des graphes, théorie des jeux...).

La convexité se retrouve dans des contextes très divers ; mais le cas fondamental, le seul qui sera considéré ici, est celui des espaces vectoriels sur le corps **R** des nombres réels.



A. Ensembles convexes

Un sous-ensemble C d'un espace vectoriel réel E est dit *convexe* si, pour tout couple

de points quelconques de C, le segment qui a pour extrémités ces deux points est entièrement contenu dans C. Par exemple, un cube est convexe, mais sa surface ne l'est pas, car elle ne contient le segment d'extrémités x et y que si x et y appartiennent à la même face. Les ensembles convexes interviennent dans de nombreux domaines des mathématiques et il est souvent possible, en pareil cas, d'obtenir d'intéressants résultats en ne faisant appel qu'à des arguments « géométriques » relativement élémentaires.

Minkowski (1864-1909) fut le premier à étudier systématiquement les ensembles convexes et ses œuvres contiennent la plupart des idées importantes utilisées pour ce sujet. Les premiers développements se limitaient aux espaces vectoriels de dimension finie et l'objet principal de ces études était de résoudre des problèmes de nature quantitative ; depuis 1940, les aspects combinatoires et qualitatifs ont bénéficié d'une plus grande attention. Après quelques préliminaires généraux, on traitera d'abord les aspects quantitatifs et combinatoires, en se limitant au cas où l'espace est de dimension finie ; on abordera ensuite les aspects qualitatifs de la théorie et ses applications à l'analyse fonctionnelle.

Un des aspects les plus fascinants de la théorie des ensembles convexes est le grand nombre de problèmes très faciles et intuitifs à formuler que l'on ne sait pas toujours résoudre.

1. Propriétés générales

Définitions

Soit x et y, deux points distincts d'un espace vectoriel réel E (cf. algèbre LINÉAIRE ET MULTILINÉAIRE). Par analogie avec le cas de l'espace usuel $\mathbf{R}^3$ (représentation paramétrique de la droite définie par deux points), on appelle *droite joignant x et y* l'ensemble des points de E de la forme :

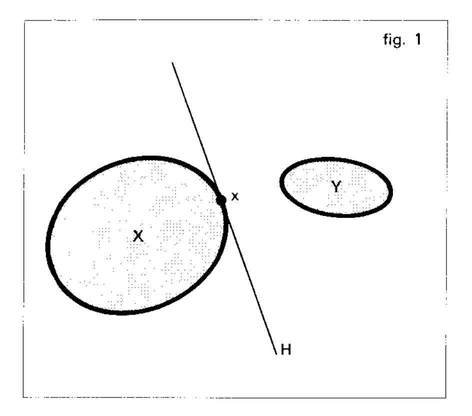
$$(1-\lambda)x + \lambda y,$$

où λ est un nombre réel quelconque. Les points tels que $\lambda \geqslant 0$ constituent la *demi-droite xy d'origine x* ; les points tels que $0 \leqslant \lambda \leqslant 1$ constituent le *segment* $[x, y]$ *d'extrémités* x et y.

Par définition, on appelle *sous-variété linéaire* de E tout sous-ensemble de E qui contient toute droite joignant deux quelconques de ses points ; par exemple, dans l'espace usuel $\mathbf{R}^3$, les sous-variétés linéaires sont : l'ensemble vide, les ensembles réduits à un point, les droites, les plans et l'espace $\mathbf{R}^3$ tout entier. Toute sous-variété linéaire V de E est la translatée d'un sous-espace vectoriel de E, c'est-à-dire l'ensemble des points de la forme $a + x$, où a est un élément fixé de V et où x parcourt un sous-espace vectoriel F de E ; si F est de dimension finie p, on dit que V est de dimension p.

Par analogie avec le cas des plans dans $\mathbf{R}^3$, on appelle *hyperplan* de E toute sous-variété linéaire qui n'est contenue strictement dans aucune autre variété linéaire que E lui-même ; par exemple, les hyperplans de $\mathbf{R}^n$ sont les variétés linéaires de dimension $n - 1$. Le complémentaire (ensembliste) d'un hyperplan H est la réunion (ensembliste) de deux ensembles convexes disjoints appelés les *demi-espaces ouverts* limités par H ; leurs réunions avec H s'appellent les *demi-espaces fermés* limités par H. On dit que deux ensembles X et Y sont *séparés* par H si l'un est contenu dans un de ces deux demi-espaces fermés et l'autre dans l'autre demi-espace ; on dit que H est un *hyperplan d'appui* de X au

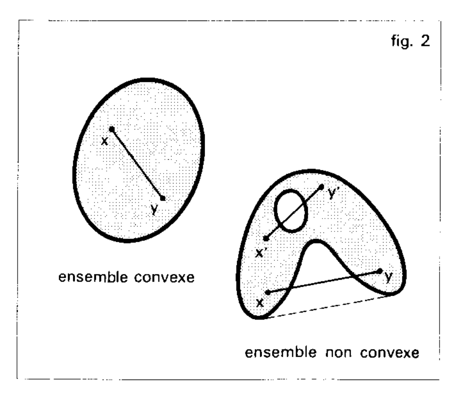
point x si x appartient à X et si X et x sont séparés par H. La figure 1 donne un



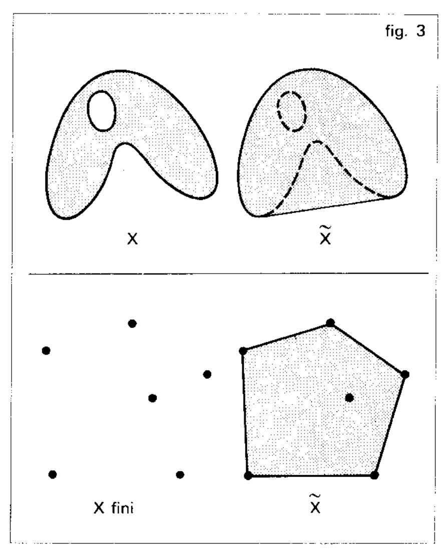
exemple d'un hyperplan H d'appui de X en x, séparant X et Y (ici $E = \mathbf{R}^2$, et H est une droite).

Ensembles convexes

Un sous-ensemble C de E est dit *convexe* si pour tout couple x, y de points distincts de C, le segment $[x, y]$ est entièrement contenu dans C (fig. 2). Il est clair que toute inter-



section d'ensembles convexes est encore un ensemble convexe. Par suite, si X est un sous-ensemble quelconque de E, l'intersection $\tilde{X}$ de tous les ensembles convexes contenant X (il y en a toujours au moins un, E lui-même) est un ensemble convexe contenant X qui possède la propriété d'être contenu dans tout ensemble convexe contenant X ; $\tilde{X}$ s'appelle l'*enveloppe convexe* de X (fig. 3). On peut aussi définir $\tilde{X}$ comme



l'ensemble de toutes les *combinaisons convexes* des points de X, c'est-à-dire l'ensemble des points de la forme :

$$\sum_{i}^{k} \lambda_i x_i,$$

où les x_i sont des points de X et où les λ_i sont des nombres réels positifs quelconques de somme 1. On appelle *enveloppe convexe fermée* de X l'adhérence de $\tilde{X}$.

Les enveloppes convexes des sous-ensembles *finis* de $\mathbf{R}^n$ sont appelées des polytopes (cf. *infra*). En particulier, on appelle *simplexe de dimension n* l'enveloppe convexe de $(n + 1)$ points distincts de $\mathbf{R}^n$ non situés dans un même hyperplan (segment pour $n = 1$, triangle pour $n = 2$, tétraèdre pour $n = 3$, etc.) ; cette notion joue un rôle fondamental en topologie algébrique.

Un point x d'un ensemble convexe C est appelé un *point extrémal* si son complémentaire dans C est convexe, c'est-à-dire si x n'appartient à l'intérieur d'aucun segment contenu dans C. Une *face* F de C est un sous-ensemble convexe de C tel qu'aucun segment contenu dans C ne traverse F, c'est-à-dire que tout segment $[x, y]$ dont les extrémités appartiennent à C et dont un point différent des extrémités appartient à F est entièrement contenu dans F. Ainsi, le cube a six 2-faces, douze 1-faces (les arêtes) et huit 0-faces (les sommets) ; de plus, le cube est l'enveloppe convexe de ses points extrémaux, qui sont ici les sommets.

Les termes corps et cônes sont utilisés de manière assez différente par plusieurs auteurs. Nous adopterons les définitions suivantes : Un *corps* est un sous-ensemble borné de $\mathbf{R}^n$ d'intérieur non vide (c'est-à-dire contenant au moins une boule de rayon > 0) et un *cône* est un sous-ensemble d'un espace affine réel E, qui est réunion d'une famille de demi-droites de même origine.

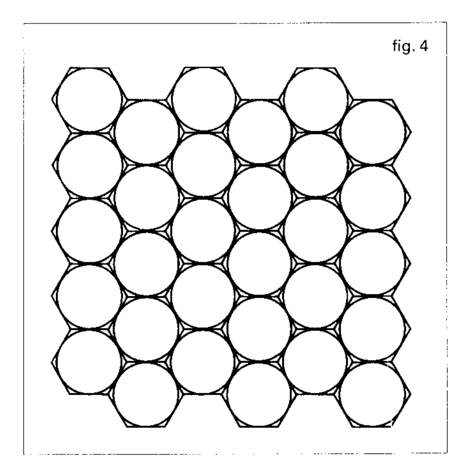
2. Aspects quantitatifs

Géométrie des nombres

Ce sont des recherches de théorie des nombres qui furent à l'origine des premiers travaux de Minkowski. Nous renvoyons, pour cet aspect, à l'article approximations DIOPHANTIENNES, en nous contentant de rappeler ici l'énoncé du célèbre théorème de Minkowski : Si C est un sous-ensemble convexe de $\mathbf{R}^n$, symétrique par rapport à l'origine, et de volume $V(C) \geqslant 2^n$, alors C contient au moins un point dont toutes les coordonnées sont des nombres entiers.

Empilements

Un empilement est un arrangement de corps convexes tel qu'aucune paire de ces corps n'ait de point intérieur en commun. Indépendamment de l'intérêt que les problèmes d'empilements ont en eux-mêmes, ce genre de problèmes se retrouve également en théorie des nombres, en théorie de l'information, en cristallographie, en botanique, en construction des réacteurs nucléaires, etc. Très souvent, on cherche à réaliser un empilement de densité maximum. Ainsi, l'empilement de densité maximum de cercles égaux dans le plan est obtenu en décomposant le plan en hexagones réguliers égaux et en inscrivant dans chacun de ces hexagones le cercle de diamètre maximum ; ainsi, chaque cercle en touche exactement six autres (fig. 4).


fig. 4

Dans le cas de l'espace à trois dimensions, on a conjecturé que l'empilement de densité maximum de *boules* égales est obtenu par une construction due à Kepler : on commence par diviser $\mathbf{R}^3$ en un échiquier à trois dimensions, où les cubes sont coloriés alternativement en blanc et en noir ; on construit ensuite des boules centrées en chacun des centres des cubes

noirs et tangentes à chacune des douze arêtes du cube ; de cette façon, chacune des boules en touche exactement douze autres. Cette conjecture n'a été démontrée que pour des empilements assez réguliers.

Inégalités

Il y a toute une série de résultats quantitatifs relatifs au volume, à la surface, au diamètre, etc., d'un corps convexe. Par exemple, l'*inégalité isopérimétrique* exprime que la surface S et le volume V d'un corps convexe C de $\mathbf{R}^n$ vérifient l'inégalité :

$$S^n \geqslant n\omega V^{n-1},$$

où ω est le volume de la boule unité de $\mathbf{R}^n$; de plus, cette inégalité est une égalité si et seulement si C est une boule, c'est-à-dire que, parmi les corps convexes de volume donné, les boules constituent ceux dont la surface est minimum. De plus, tout corps convexe C de $\mathbf{R}^n$ est contenu dans une boule de rayon minimum r, et cette boule est unique ; l'inégalité de Jung affirme que si on désigne par d le diamètre du corps C (c'est la borne supérieure des longueurs des segments dont les extrémités appartiennent à C), on a :

$$r \leqslant (n/(2n+2))^{1/2} d\,;$$

cette inégalité devient une égalité si et seulement si C est un simplexe régulier de $n+1$ sommets dans $\mathbf{R}^n$. Un théorème de Loewner affirme que tout corps de $\mathbf{R}^n$ est contenu dans un ellipsoïde de volume minimum ; cet ellipsoïde joue un rôle important dans la théorie des modèles expérimentaux.

Volumes mixtes

Soit $C_1, C_2, ..., C_k$ des corps convexes de $\mathbf{R}^n$, et $\lambda_1, \lambda_2, ..., \lambda_k$ des nombres réels positifs ; l'ensemble des points de la forme :

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k,$$

où x_i parcourt C_i pour tout i, est un corps convexe C, que nous désignerons par :

$$\lambda_1 C_1 + ... + \lambda_k C_k\,;$$

lorsque $C_1, ..., C_k$ sont fixés, le volume de C s'exprime par un polynôme homogène de degré n en les variables $\lambda_1, ..., \lambda_k$. Certains des problèmes les plus fondamentaux de la théorie quantitative des corps convexes sont liés à l'étude des coefficients de ces polynômes, appelés *volumes mixtes* de $C_1, ..., C_k$. L'outil de base, dans l'étude des volumes mixtes, est le théorème de Brunn-Minkowski, qui affirme que, pour tout λ compris entre 0 et 1, on a :

$$[V((1-\lambda)C_1 + \lambda C_2)]^{1/n} \geqslant (1-\lambda)[V(C_1)]^{1/n} + \lambda[V(C_2)]^{1/n}$$

c'est-à-dire que la racine n-ième du volume est une fonction concave de λ. Les inégalités pour les volumes mixtes engendrent de nombreuses inégalités d'intérêt géométrique, en particulier l'inégalité isopérimétrique.

Corps de largeur constante

Un corps convexe C de $\mathbf{R}^n$ est dit *de largeur constante* b si la distance entre n'importe quelle paire d'hyperplans d'appui de C parallèles est constante, égale à b. Contrairement à l'intuition, un tel corps n'est pas nécessairement circulaire (dans le plan) ou sphérique. La définition précédente équivaut à la suivante : C a pour diamètre b et tout ensemble réunion de C et d'un point quelconque de $\mathbf{R}^n$ n'appartenant pas à C est de diamètre supérieur à b. Par suite, tout ensemble de

$\mathbf{R}^n$ de diamètre $\leqslant b$ est contenu dans un corps convexe de largeur constante b. Cette propriété permet, par exemple, de ramener le problème suivant de Borsuk au cas où X est de largeur constante : Peut-on recouvrir tout ensemble X de diamètre 1 de $\mathbf{R}^n$ par $(n + 1)$ ensembles de diamètres < 1 ? La réponse est affirmative si $n \leqslant 3$, mais inconnue pour $n > 3$.

Euler (en 1778) et de nombreux mathématiciens après lui ont étudié les corps convexes de largeur constante dans le plan $(n = 2)$; les propriétés trouvées ont été utilisées en cinématique et même pour construire un foret utilisé pour forer des trous carrés : un corps possédant ces propriétés peut être placé dans un carré de telle sorte qu'en les tournant il reste perpétuellement en contact avec les quatre côtés de la boîte.

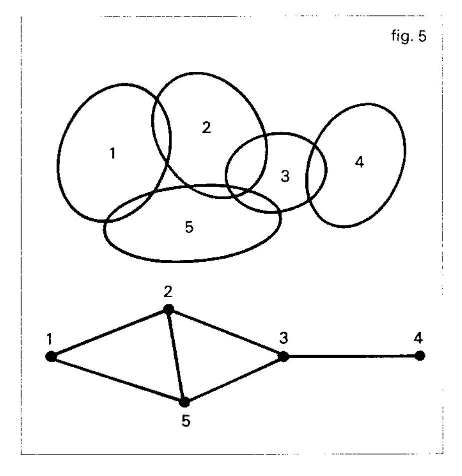
3. Aspects combinatoires

Intersections

Une partie des problèmes combinatoires est reliée à l'étude des intersections d'ensembles convexes qui sont toujours convexes, comme on l'a vu ci-dessus (l'ensemble vide est, par définition, convexe).

D'après un théorème démontré par Helly, l'intersection C d'une famille de convexes de $\mathbf{R}^n$, telle que l'intersection de toute sous-famille de $(n + 1)$ de ces ensembles soit non vide, est non vide si l'une ou l'autre des hypothèses suivantes est réalisée : la famille est finie ou chacun des convexes de la famille est fermé et borné (c'est-à-dire compact). Ce théorème admet de nombreuses généralisations et applications.

L'étude des propriétés des intersections d'ensembles convexes est facilitée par la notion de graphe d'intersection, qui est utilisée dans des domaines aussi variés que la génétique moléculaire, la psychologie et l'écologie. Pour toute famille d'ensembles, on appelle *graphe d'intersection* un graphe abstrait où chaque ensemble correspond à un sommet du graphe et où chaque intersection non vide est représentée par un arc réunissant les sommets correspondants ; la figure 5 donne un exemple d'une famille



d'ensembles convexes et de leur graphe d'intersection. Tout graphe ayant un nombre fini d'éléments est un graphe d'intersection d'ensembles convexes de $\mathbf{R}^3$, mais pas nécessairement un graphe d'intersection d'ensembles convexes de $\mathbf{R}^2$ ou de $\mathbf{R}$. Un *graphe d'intervalles* est un graphe d'intersection d'une famille finie d'ensembles convexes de $\mathbf{R}$ (ce sont des intervalles) ; on peut caractériser ces graphes d'intervalles, mais le problème correspondant pour le plan n'est pas résolu.

Étude des enveloppes convexes

Une autre série de problèmes combinatoires est la recherche de formes algébriques pour la représentation des enveloppes convexes. Voici, dans cet ordre d'idées, un

théorème très simple et très utile, dû à Carathéodory : Si X est un sous-ensemble de $\mathbf{R}^n$ et u un point de l'enveloppe convexe de X, alors u appartient à l'enveloppe convexe d'un sous-ensemble fini Y de X contenant au plus $(n + 1)$ points. Par exemple, si X est un sous-ensemble du plan et si x appartient à l'enveloppe convexe de X, alors x appartient soit à X, soit à un segment ayant pour extrémités deux points de X, soit à un triangle ayant pour sommet trois points de X. Le théorème de Carathéodory a de nombreuses applications ; on en déduit, par exemple, que l'enveloppe convexe d'un ensemble fermé borné de $\mathbf{R}^n$ est un ensemble fermé borné.

Polyèdres

Parmi tous les problèmes combinatoires que l'on rencontre dans la théorie de la convexité, celui qui est le plus ancien et qui a été étudié de la manière la plus approfondie est la structure des faces des polyèdres convexes. Nous appellerons *polyèdre* tout sous-ensemble de $\mathbf{R}^n$ intersection d'un nombre fini de demi-espaces fermés, en réservant la dénomination de *polytope*, aux polyèdres *bornés* ; un lemme, dû à Farkas, montre qu'un ensemble est un polytope si et seulement s'il est l'enveloppe convexe d'un nombre *fini* de points. On peut caractériser de même les cônes polyédraux comme enveloppes convexes d'un nombre fini de demi-droites de même origine. De manière générale, un sous-ensemble P de $\mathbf{R}^n$ est un polyèdre si et seulement s'il satisfait à une des conditions équivalentes suivantes :

a) P est un ensemble convexe fermé qui a un nombre fini de faces ;

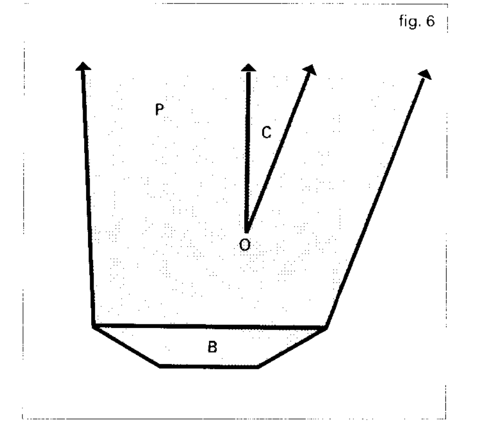
b) P est l'enveloppe d'un nombre fini de points et de demi-droites ;

c) P est la somme vectorielle B + C d'un polytope B et d'un cône polyédral C, c'est-à-dire P est l'ensemble des points de la forme :

$$x + y, \quad x \in \mathbf{B}, \quad y \in \mathbf{C}.$$

d) P est l'enveloppe convexe fermée de la réunion d'un polytope B et du translaté d'un cône polyédral. (On a représenté, sur la figure 6, un polyèdre P de dimension 2,

fig. 6



ainsi que le polytope B et le cône C mentionnés ci-dessus.)

Le premier résultat, dans l'étude combinatoire des polytopes est le théorème d'Euler (1752), qui affirme que si v, e, f sont respectivement le nombre de sommets, d'arêtes et de faces d'un polytope *de dimension 3*, on a :

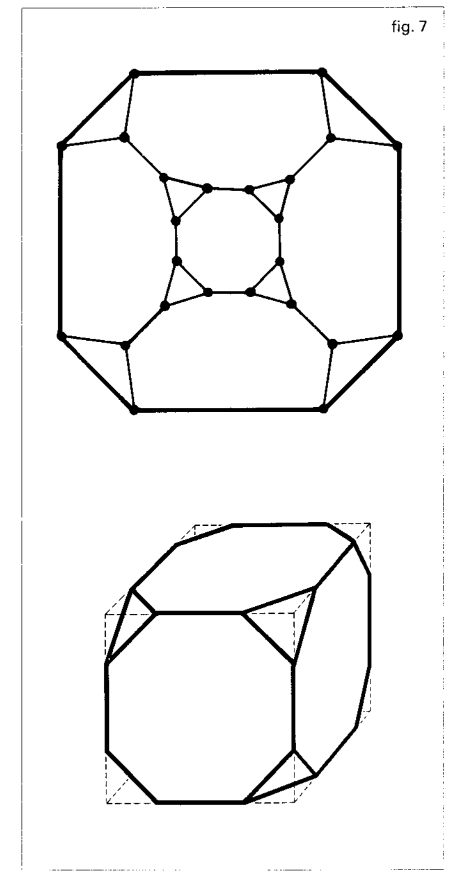
$$v - e + f = 2$$

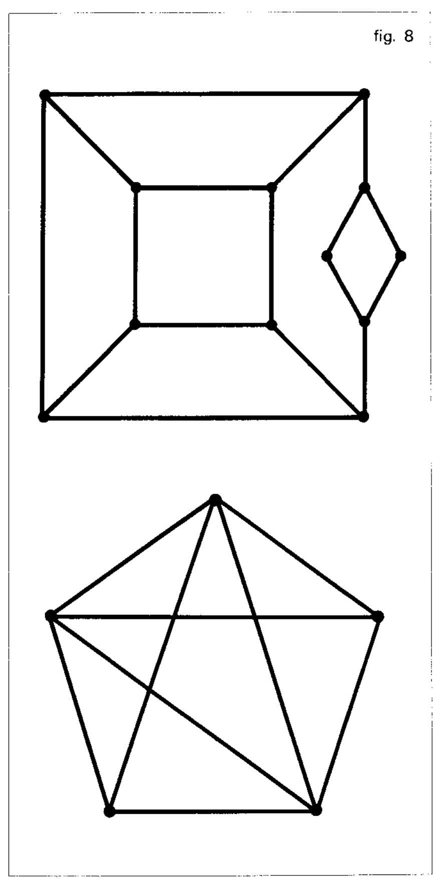
(on appelle ici sommets les points extrémaux du polyèdre). Poincaré et Schläfli ont généralisé ce théorème aux polytopes de dimension n : si f_i(P) est le nombre de faces de P de dimensions i, on a la relation :

$$\sum_{i=0}^{n-1} (-1)^i f_i(\mathrm{P}) = 1 - (-1)^n.$$

En 1934, Steinitz est parvenu à *caractériser* les graphes des polytopes de dimen-

sion 3 (on appelle graphe d'un polytope la structure combinatoire déterminée par les sommets et les arêtes) : le graphe d'un polytope de dimension 3 est équivalent à un graphe planaire (qui peut se dessiner dans le plan sans qu'il y ait aucune intersection d'arcs) connexe de degré 3 (c'est-à-dire qui ne peut être séparé en deux parties disjointes en enlevant moins de trois sommets). Par exemple, le graphe de la figure 7 est le graphe du polytope obtenu en tronquant chacun des sommets d'un cube ; les deux graphes de la figure 8 ne sont pas des graphes de polytopes de dimension 3, car le premier n'est pas connexe de degré 3 et le second n'est pas semblable à un graphe planaire (c'est le graphe d'un simplexe de dimension 4). On a découvert de nombreuses propriétés des graphes des polytopes de dimension n (par exemple, ils sont connexes de degré n), mais on ne connaît à ce jour aucune caractérisation combinatoire des graphes des polytopes de dimension > 3.


fig. 7


fig. 8

De nombreux problèmes de programmation linéaire et d'optimisation reviennent à trouver les points d'un polyèdre P où une fonction linéaire atteint son minimum ; on montre que, si P est borné, le

minimum est alors atteint en un des sommets de P et certains procédés de résolution de programmes linéaires reposent sur cette propriété. On est donc amené à estimer le nombre de sommets d'un polytope P en fonction de sa dimension et du nombre de faces de dimension $(n-1)$. Pour $2 \leqslant n < k$, désignons par $\mu(n, k)$ le nombre maximum de sommets que possède un polyèdre de dimension n qui a k faces de dimension $(n-1)$ et par $\varphi(n, k)$ la somme des coefficients binomiaux :

$$\varphi(n, k) = C_p^{k-n} + C_q^{k-n},$$

pour $p = k - [(n+1)/2], q = k - [(n+2)/2]$, en désignant par $[r]$ la partie entière de r, c'est-à-dire le plus grand entier $\leqslant r$; avec ces notations, on a alors :

$$\mu(n, k) \geqslant \varphi(n, k),$$

et cette inégalité devient en fait une *égalité* si $n \leqslant 8$, ou $k \leqslant n+3$, ou $k \geqslant n^2/4$ (on ne sait s'il y a toujours égalité). L'étude de la fonction μ montre, par exemple, le fait remarquable que pour tout $k > n$, il y a un polytope de dimension n à k sommets tel que chaque groupe de $[n/2]$ sommets détermine une face : ainsi, il y a des polytopes de dimension 4 avec un nombre quelconque de sommets où chaque paire de sommets est reliée par une arête du polytope.

4. Aspects qualitatifs

Contrairement à ce qui précède, les espaces considérés ici sont quelconques, et non nécessairement de dimension finie.

La convexité intervient de manière essentielle dans les espaces vectoriels de l'analyse : espaces vectoriels normés, ou plus généralement espaces vectoriels topologiques localement convexes, c'est-à-dire où tout point a un système fondamental de voisinages convexes ; on se limitera ici à de rapides indications, en renvoyant pour les définitions aux articles espaces vectoriels NORMÉS et espaces vectoriels TOPOLOGIQUES.

Espaces normés

On soulignera seulement le rôle de la convexité. Rappelons qu'une norme sur un espace vectoriel E (qui sera ici réel) est une fonction p à valeurs positives définie dans E telle que :

a) $p(x) = 0$ si et seulement si $x = 0$;

b) $p(\lambda x) = |\lambda| p(x)$ pour $x \in E$ et $\lambda \in R$;

c) $p(x+y) \leqslant p(x) + p(y)$ (sous-additivité).

Une norme est souvent notée $\|.\|$.

L'étude des espaces vectoriels normés est, d'une certaine façon, l'équivalent de l'étude d'une classe d'ensembles convexes ; le lien est établi par la fonction de jauge. Si $\|.\|$ est une norme sur E, on appelle *boule unité* (resp. *sphère unité*) pour cette norme l'ensemble des $x \in E$ tels que $\|x\| \leqslant$ (resp. $\|x\| = 1$). Les propriétés de la norme entraînent que la boule unité U est un ensemble convexe dont l'intersection avec toute droite passant par O est un segment symétrique $[-x, x]$ (non réduit au point O). Réciproquement, soit U un ensemble convexe satisfaisant à ces propriétés ; pour tout $x \neq 0$, désignons par $p(x)$ le plus petit entier positif λ tel que $x/\lambda \in U$ et posons $p(0) = 0$ (la fonction p est appelée la *jauge* de l'ensemble U). On vérifie facilement que la jauge de U est une norme sur E pour laquelle U est la boule unité (en fait, la jauge d'un ensemble peut se définir sous des hypothèses beaucoup plus générales). Ainsi toutes les propriétés d'un espace normé peuvent être décrites uniquement en fonction de sa boule ou

de sa sphère unité. Par exemple, la *convexité stricte* d'un espace vectoriel normé, caractérisée par l'inégalité stricte $\|x+y\| < \|x\| + \|y\|$ lorsque x et y ne sont pas sur une même demi-droite d'origine O, équivaut au fait que sa sphère unité ne contient aucun segment.

On appelle *espace de Minkowski* tout espace vectoriel normé de dimension finie. Par exemple, pour tout $r \geqslant 1$:

$$\|x\|_r = \left(\sum_{i=1}^{n} |x_i|^r\right)^{1/r},$$

où $x = (x_1, ..., x_n) \in \mathbf{R}^n$, est une norme sur $\mathbf{R}^n$; pour $r = 2$, on retrouve la norme euclidienne usuelle. Une autre norme importante sur $\mathbf{R}^n$ est la norme :

$$\|x\|_\infty = \sup_{1 \leqslant i \leqslant n} |x_i|,$$

$x = (x_1, ..., x_n) \in \mathbf{R}^n$, pour laquelle la boule unité est l'hypercube de $\mathbf{R}^n$.

Tous les problèmes quantitatifs indiqués ci-dessus ont fait l'objet d'études analogues pour les espaces de Minkowski : inégalité isopérimétrique, inégalité de Jung, etc.

Théorèmes de séparation

En analyse fonctionnelle, en théorie des jeux, en intégration et même dans certains problèmes relatifs aux graphes coloriés en théorie des graphes, on utilise des théorèmes de séparation et de support.

Les théorèmes de séparation établissent les conditions sous lesquelles on peut séparer (au sens du chapitre 1) deux sous-ensembles convexes disjoints X et Y d'un espace vectoriel topologique E. Pour que cela soit possible, il suffit, par exemple, que l'une des conditions suivantes soit réalisée :

a) E est de dimension finie ;

b) un des deux ensembles a un intérieur non vide ;

c) un des deux ensembles est fermé, l'autre compact, et E est localement convexe.

Dans chacun de ces cas, on peut choisir l'hyperplan de séparation fermé, c'est-à-dire défini comme l'ensemble des zéros d'une fonction affine *continue.*

Remarquant que l'intérieur d'un ensemble convexe est convexe, on en déduit que si A est un ensemble convexe d'intérieur non vide et A une sous-variété linéaire de E ne rencontrant pas l'intérieur de C, alors il existe un hyperplan qui contient A et qui sépare A de C (forme « géométrique » du théorème de Hahn-Banach). En particulier, si C est un ensemble convexe d'intérieur non vide, C admet un hyperplan d'appui en chaque point de sa frontière (théorème de Mazur). On ne peut pas étendre ce théorème au cas où C n'a pas de point intérieur, mais on peut montrer que, dans certains cas (C fermé dans un espace de Banach, ou C compact pour la topologie faible d'un espace localement convexe), les points de la frontière de C, où C admet un hyperplan d'appui, forment un sous-ensemble dense de la frontière de C.

Soit A une sous-variété linéaire d'un espace vectoriel E, p une fonction convexe dans E (cf. CONVEXITÉ - Fonctions convexes), et f une fonction affine définie dans A telle que $f(x) \leqslant p(x)$ en tout point x de A ; alors il existe une fonction affine g définie dans E qui prolonge f (c'est-à-dire que $g(x) = f(x)$ pour tout $x \in$ A) telle que l'on ait $g(x) \leqslant p(x)$ en tout point $x \in$ E (théorème de Hahn-Banach). Ce résultat, essentiel en analyse fonctionnelle, équivaut en fait aux théorèmes énoncés ci-dessus : il résulte, par exemple, du théorème de séparation en prenant pour ensembles

le graphe de la fonction f et l'épigraphe de la fonction p. D'autre part, ce théorème de Hahn-Banach entraîne que si C est un ensemble convexe contenant l'origine comme point intérieur et si u est un point frontière de C, alors C admet un hyperplan d'appui en u (on retrouve le théorème de Mazur ; il faut prendre comme fonction convexe p la jauge de C).

Points extrémaux

Les fonctions convexes et concaves présentent des propriétés très utiles dans les problèmes d'optimisation. Par exemple, soit f une fonction convexe définie dans un domaine convexe D d'un espace vectoriel topologique localement convexe E ; alors tout minimum *local* de f dans D est un minimum *global*, c'est-à-dire que si un point x_0 de D possède un voisinage U tel que $f(x) \geqslant f(x_0)$ pour tout $x \in U \cap D$, alors cette inégalité est valable pour tout $x \in D$. Si f est concave et D compact, alors f atteint son minimum en un point extrémal de D ; cette propriété justifie l'étude des points extrémaux en analyse fonctionnelle.

Le théorème de Krein-Milman attire l'attention sur d'autres propriétés importantes des points extrémaux. Ce théorème établit que si C est un sous-ensemble compact convexe d'un espace localement convexe et si X est un sous-ensemble de C, alors C est l'enveloppe convexe fermée de X si et seulement si l'adhérence de X contient tous les points extrémaux de C ; ainsi, l'adhérence de l'ensemble des points extrémaux de C est le plus petit sous-ensemble X de C tel que chaque point de C soit adhérent à l'ensemble des combinaisons convexes (cf. chap. 1) de points de X. On peut généraliser ce théorème au cas où C n'est pas compact moyennant des hypothèses supplémentaires ; par exemple, si C est un convexe fermé de $\mathbf{R}^n$ ne contenant aucune droite, alors C est l'enveloppe convexe de ses points extrémaux et de ses demi-droites extrémales (il n'est pas nécessaire de prendre l'enveloppe convexe fermée ; par définition, une demi-droite extrémale de C est une demi-droite qui n'est « traversée » par aucun segment contenu dans C).

Voici une application du théorème de Krein-Milman à l'analyse fonctionnelle. Soit C un sous-ensemble compact convexe d'un espace localement convexe E et soit X l'ensemble de tous les points extrémaux ; alors, d'après le théorème de Krein-Milman, pour tout point p de C, il existe une mesure positive μ sur l'adhérence $\bar{X}$ de X telle que $\mu(X) = 1$ et telle que :

$$f(p) = \int_{\bar{X}} f(x)\, d\mu(x)$$

pour toute forme linéaire continue f (p est le « barycentre » de la mesure μ). Ce résultat entraîne de nombreux théorèmes de représentation intégrale en analyse.

Théorèmes de point fixe

Indiquons rapidement, pour conclure, deux théorèmes de point fixe pour les ensembles convexes.

Le théorème de Brouwer-Schauder-Tychonov montre que si C est un compact convexe d'un espace localement convexe et f une application continue de C dans lui-même, il existe au moins un point p de C tel que $f(p) = p$. Ce théorème permet d'obtenir des théorèmes d'existence pour les solutions des équations différentielles et intégrales, les théorèmes du minimax en théorie des jeux et de nombreuses propriétés des ensembles convexes. Par exemple un compact C de $\mathbf{R}^n$ ou d'un espace de Hilbert H est convexe si pour chaque point x de l'espace, il y a un point unique de C

qui minimise la distance entre x et C (pour $\mathbf{R}^n$, il suffit que C soit fermé ; dans le cas d'un espace de Hilbert quelconque le problème n'est pas résolu).

Le théorème de Markov-Kakutani affirme que si C est un compact convexe d'un espace vectoriel topologique et si Φ est une famille commutative de transformations affines continues de C dans C, alors il y a au moins un point p de C tel que $f(p) = p$ pour toute fonction de Φ. Ce théorème sert à établir l'existence de mesures invariantes pour les groupes commutatifs et à construire des mesures qui généralisent la mesure de Lebesgue pour les ensembles bornés de $\mathbf{R}^n$.

VICTOR KLEE

Bibliographie

M. BERGER, *Convexes et polytopes, polyèdres réguliers, aires et volumes*, C.E.D.I.C., Paris, 1978 / T. BONNESEN & W. FENCHEL, *Theorie der konvexen Körper*, Springer, Berlin, 1974 / N. BOURBAKI, *Espaces vectoriels topologiques*, Masson, 1981 / J. C. CONWAY & M. J. SLOANE, *Spere Packings, Lattices and Groups*, Springer-Verlag, New York, 1987 / H. T. CROFT, K. J. FALCONER & R. K. GUY, *Unsolved Problems in Geometry*, *ibid.*, 1991 / B. GRÜNBAUM, *Convex Polytopes*, New York, 1967 / L. JOLY, *Les Polyèdres : réguliers, semi-réguliers et composés*, A. Blanchard, 1979 / V. KLEE dir., *Convexity*, Proceedings of 7th Symposium of the American Mathematical Society held at the University of Washington, Seattle, 1963, repr. A.M.S., Providence, 1979 / S. LANG, *Analyse réelle*, Interéditions, 1977 / H. MOULIN, *La Convexité dans les mathématiques de la décision*, Hermann, 1979 / P. & S. PEARCE, *Polyhedra Primer*, Dale Seymour Publ., Palo Alto (Calif.), 1978.

B. Fonctions convexes

L'étude des fonctions convexes a permis de fournir un cadre dans lequel peut se résoudre toute une classe de problèmes d'analyse fonctionnelle non linéaire ; les problèmes ainsi abordés sont des questions d'optimisation provenant de divers domaines : la mécanique, l'économie, les équations aux dérivées partielles, l'analyse numérique. Compte tenu de la difficulté d'aborder de manière un peu générale les problèmes non linéaires, c'est là un rôle très important qui a motivé le développement autonome de la théorie.

Les travaux de W. Fenchel, de T. Rockafellar, de J.-J. Moreau ont développé les outils de base de l'*analyse convexe* notamment la notion de fonctions convexes conjuguées et la notion de sous-différentiel qui sert de produit de remplacement pour les fonctions convexes non différentiables.

Nous renvoyons à la partie **A** ci-dessus - Ensembles convexes, pour tout ce qui concerne les résultats généraux sur les convexes.

1. Les fonctions convexes

Soit E un espace vectoriel sur $\mathbf{R}$, C une partie convexe de E et f une fonction définie sur E à valeurs dans $\overline{\mathbf{R}}$ (c'est-à-dire prenant éventuellement les valeurs $\pm\infty$). L'*épigraphe* de f, noté épi(f), est l'ensemble des couples (x, a) de $C \times \mathbf{R}$ tels que $f(x) \leqslant a$. La fonction f sera dite *convexe* si son épigraphe est une partie convexe de $E \times \mathbf{R}$.

On obtient immédiatement une interprétation analytique de cette définition : La fonction f est convexe si et seulement si, pour tout réel λ de l'intervalle $[0, 1]$, on a :

$$(1) \quad f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y)$$

pour tous les couples (x, y) d'éléments de C ne vérifiant pas $f(x) = -f(y) = \pm\infty$ (auquel cas le second membre de l'inégalité (1) n'est pas défini). En raisonnant par récurrence, on prouve que, si $\lambda_1, \lambda_2, ..., \lambda_n$

sont des réels positifs dont la somme est 1, on a :

$$(2) \qquad f(\sum_{i=1}^{n} \lambda_i x_i) \leqslant \sum_{i=1}^{n} \lambda_i f(x_i)$$

chaque fois que le second membre de l'inégalité (2) a un sens.

La possibilité pour la fonction f de prendre la valeur $+\infty$ permet de ne considérer que des fonctions convexes définies sur E tout entier ; en effet, si on prolonge la fonction f définie sur C en la fonction $\tilde{f}$ définie sur E en posant $\tilde{f}(x) = +\infty$ si $x \notin C$, les fonctions f et $\tilde{f}$ ont alors le même épigraphe et donc f est convexe si et seulement si $\tilde{f}$ est convexe. Désormais, nous ne considérerons donc que des fonctions définies sur E tout entier. Cela nous conduit à définir le *domaine effectif* de f, noté dom (f) :

$$\text{dom}(f) = \{x \in E\,; f(x) < +\infty\}.$$

Le domaine effectif de f est la projection sur E de l'épigraphe de f ; c'est une partie convexe de E.

La valeur $-\infty$ peut se présenter dans certains cas particuliers ; nous ne l'éliminons pas a priori ; néammoins, nous introduisons la terminologie suivante : La fonction convexe f est propre si son domaine effectif est non vide et si elle ne prend jamais la valeur $-\infty$; la restriction de f à dom (f) est alors une fonction à valeurs dans **R** (cf. la partie **A** ci-dessus - Ensembles convexes). Une fonction deux fois continûment différentiable sur un ouvert convexe C de $\mathbf{R}^n$ à valeurs réelles est convexe si et seulement si la matrice hessienne :

$$\left(\frac{\partial f}{\partial x_i \partial x_j}(x)\right)_{i,j}$$

est, en tout point x de C, la matrice d'une forme quadratique positive.

2. Cas de la dimension 1

L'exemple des fonctions convexes définies sur **R** est instructif pour l'étude ultérieure des fonctions convexes définies sur $\mathbf{R}^n$, ou même sur des espaces vectoriels topologiques. En outre, ce cas a un intérêt propre pour la définition d'une classe intéressante d'espaces : les espaces d'Orlicz. Dans tout ce chapitre 2, f est une fonction convexe propre définie sur **R**.

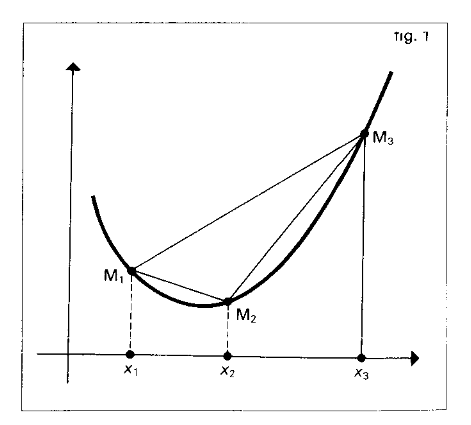
Supposons que x_1, x_2, x_3 soient dans dom (f) et vérifient $x_1 \leqslant x_2 \leqslant x_3$; en remarquant que :

$$x_2 = \frac{x_3 - x_2}{x_3 - x_1} x_1 + \frac{x_2 - x_1}{x_3 - x_1} x_3$$

et en appliquant l'inégalité (1), on obtient les inégalités :

$$(3) \qquad \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leqslant \frac{f(x_3) - f(x_1)}{x_3 - x_1} \leqslant \frac{f(x_3) - f(x_2)}{x_3 - x_2},$$

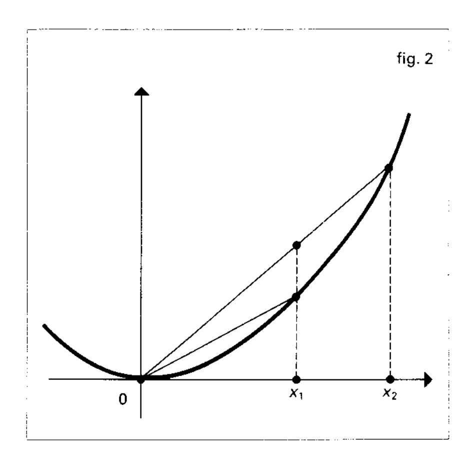
c'est-à-dire que le coefficient directeur de la droite M_1M_3 (cf. fig. 1) est compris entre



celui de la droite M_1M_2 et celui de la droite M_2M_3. En se servant de ces inégalités, on montre que f est continue sur l'intérieur de son domaine effectif.

L'utilisation des inégalités (3) permet de montrer que, en tout point intérieur à dom f, la fonction f admet une dérivée à droite $f'_d(x)$ et une dérivée à gauche $f'_g(x)$ et qu'on a, en outre, $f'_g(x) \leqslant f'_d(x)$. De plus, $f'_d(x)$ est croissante et continue à droite sur l'intérieur de dom f. Si x_0 est un point intérieur à dom f tel que $f(x_0) = 0$, pour tout x intérieur à dom f, on peut écrire :

$$f(x) = \int_{x_0}^{x} f'_d(t)\,dt.$$



Les N-fonctions

Considérons maintenant les fonctions convexes définies sur **R** à valeurs dans **R** et qui admettent une représentation de la forme :

(4) $$f(x) = \int_0^{\varphi|x|} \varphi(t)\,dt,$$

où φ est une fonction définie sur $[0, +\infty[$, croissante, continue à droite, nulle en 0, telle que :

$$\lim_{x \to +\infty} \varphi(x) = +\infty \text{ et } \varphi(x) > 0 \text{ pour } x > 0;$$

ces fonctions présentent un intérêt particulier pour la définition des espaces d'Orlicz ; ce sont les N-fonctions. Il s'agit, en fait, des fonctions convexes paires définies sur **R** à valeurs dans **R** strictement croissantes sur $[0, +\infty[$, telles que :

$$\lim_{x \to 0} \frac{f(x)}{x} = 0 \quad \text{et} \quad \lim_{x \to +\infty} \frac{f(x)}{x} = +\infty.$$

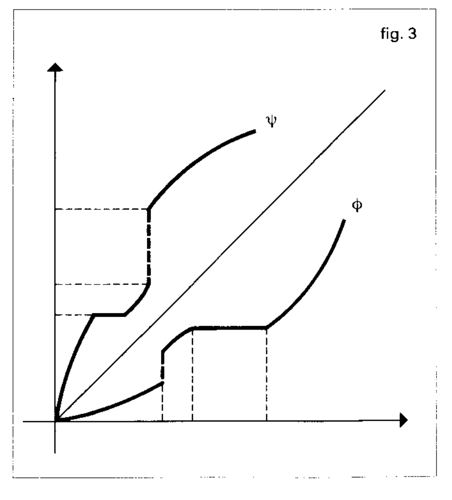
Les N-fonctions vérifient, en outre, les inégalités (cf. fig. 2) :

$$f(\alpha x) \leqslant \alpha f(x), \text{ si } x \in \mathbf{R} \text{ et } 0 \leqslant \alpha \leqslant 1,$$
$$f(\alpha x) < \alpha f(x), \text{ si } x \neq 0 \text{ et } 0 < \alpha < 1,$$
$$\frac{f(x_1)}{x_1} < \frac{f(x_2)}{x_2}, \text{ si } 0 < x_1 < x_2.$$

Soit f une N-fonction exprimée sous la forme (4) ; posons :

(5) $$\psi(x) = \sup\{u\,;\, \varphi(u) \leqslant x\}.$$

Si φ est continue strictement croissante, ψ est la fonction réciproque de φ (fig. 3).



Remarquons qu'à un intervalle sur lequel φ est constante correspond un saut de la fonction ψ et qu'à un saut de la fonction φ correspond un intervalle sur lequel ψ est constante ; si l'on rajoute aux

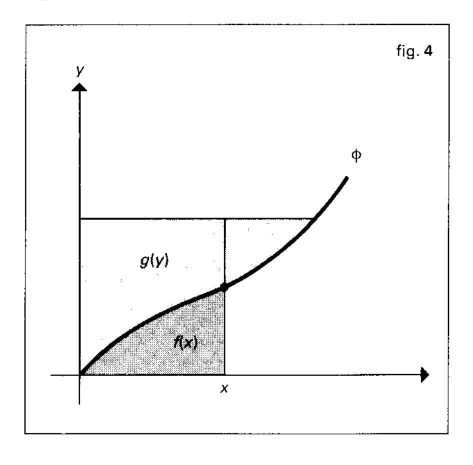
courbes représentatives de φ et de ψ les segments verticaux qui correspondent aux sauts des fonctions φ et ψ (ce sont les seules discontinuités possibles puisque ces fonctions sont monotones), on obtient des courbes symétriques par rapport à la première bissectrice. Une démarche analogue effectuée sur la fonction ψ redonne la fonction φ.

Si, maintenant, on pose :

$$g(x) = \int_0^{|x|} \psi(t)\,dt,$$

on obtient une N-fonction appelée *fonction conjuguée* de la fonction f.

L'inégalité suivante, appelée inégalité de Young, dont la signification géométrique obtenue en interprétant $f(x)$ et $g(y)$ comme des aires est suggérée sur la figure 4, a lieu :



(6) $$xy \leqslant f(x) + g(y).$$

L'égalité est atteinte lorsque $x \geqslant 0$ et $y = \varphi(x)$; si bien que l'on a :

$$|x|\varphi(|x|) = f(x) + g(\varphi(|x|))$$

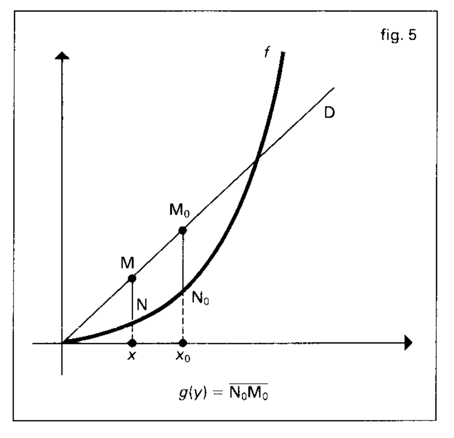
et qu'on a de même :

$$|y|\psi(|y|) = g(y) + f(\psi(|y|)).$$

Comme $g(y) \geqslant xy - f(x)$ et que l'égalité a lieu pour au moins une valeur x_0 de x, on peut dire que :

(7) $$g(y) = \max_x(xy - f(x)).$$

Cette inégalité a une interprétation géométrique simple en introduisant le point M d'abscisse x sur la droite D passant par l'origine de coefficient directeur y et le point N d'abscisse x sur la courbe représentative de g (cf. fig. 5) ;



alors :

$$g(y) = \max_x \overline{\mathbf{NM}} = \overline{\mathbf{N_0M_0}}.$$

Remarquons que, si f est dérivable, la tangente au graphe de f en N_0 est parallèle à la droite D et l'équation de cette tangente est $t(x) = xy - g(y)$. La fonction g est la transformée de Legendre de f.

On peut aussi dire que la fonction $t(x) = xy - g(y)$ est la plus grande fonction affine de coefficient directeur y qui minore f.

Si $p > 1$, la fonction $f(x) = \frac{1}{p}|x|^p$ est un exemple de N-fonction ; pour $x > 0$, on a $f'(x) = x^{p-1}$ et $(f')^{-1}(x) = x^{q-1}$, où $1/p + 1/q = 1$; par conséquent la fonction g conjuguée de f est définie par $g(x) = \frac{1}{q}|x|^q$.

Les espaces d'Orlicz

Soit f une N-fonction, notons l_f l'ensemble des suites réelles $(x_i)_{i\geqslant 0}$ telles qu'il existe $\alpha > 0$ pour lequel :

$$(8) \qquad \sum_{i=0}^{+\infty} f\left(\frac{x_i}{\alpha}\right) < +\infty.$$

l_f est un sous-espace vectoriel de l'espace des suites que l'on munit d'une norme en posant :

$$\|(x_i)_{i\geqslant 0}\| = \inf\{\alpha > 0; \sum_{i=0}^{+\infty} f\left(\frac{x_i}{\alpha}\right) \leqslant 1\}.$$

Muni de cette norme, l_f est un espace de Banach (cf. espaces vectoriels NORMÉS), appelé espace d'Orlicz de suites associé à la N-fonction f.

On peut montrer que l_f est aussi l'ensemble des suites réelles $(x_i)_{i\geqslant 0}$ telles que :

$$\sup\{|\sum_{i=0}^{+\infty} x_i y_i|; \sum_{i=0}^{+\infty} g(y_i) \leqslant 1\} < +\infty,$$

où g est la fonction conjuguée de f.

On définit alors une autre norme sur l_f en posant :

$$|||(x_i)_{i\geqslant 0}||| = \sup\{|\sum_{i=0}^{+\infty} x_i y_i|; \sum_{i=0}^{+\infty} g(y_i) \leqslant 1\}.$$

Cette norme est équivalente à la première ; plus précisément :

$$\|(x_i)_{i\geqslant 0}\| \leqslant |||(x_i)_{i\geqslant 0}||| \leqslant C\|(x_i)_{i\geqslant 0}\|.$$

Lorsque f vérifie, en outre, la condition :

(9) pour tout $\lambda > 1$, il existe $K > 0$ et $x > 0$ tels que $f(\lambda x) \leqslant K f(x)$;

ce qui se produit si et seulement si :

$$\limsup_{x\to 0^+} \frac{x\varphi(x)}{f(x)} < +\infty,$$

l'espace l_f est aussi l'espace des suites $(x_i)_{i\geqslant 0}$ tel que, pour tout $\alpha > 0$, l'inégalité (8) ait lieu ; l_f est alors un espace de Banach séparable ; son dual est isomorphe à l_g, où g est la conjuguée de f, grâce à l'isomorphisme :

$$T : l_g \to l_f^*,$$

où :

$$x^*((x_i)_{i\geqslant 0}) = \sum_{i=0}^{+\infty} y_i x_i,$$

avec :

$$(y_i)_{i\geqslant 0} \mapsto x^*.$$

L'application de ces considérations à la N-fonction $f(x) = \frac{1}{p}|x|^p$, avec $p > 1$, donne pour espace d'Orlicz de suites l'espace l_p. Une étude analogue peut être conduite dans le cadre de l'intégration, en définissant l'ensemble $L_f(K)$ des fonctions $x(t)$, définies à un ensemble de mesure nulle près (cf. INTÉGRATION ET MESURE), d'un compact K de $\mathbf{R}^n$, à valeurs dans $\mathbf{R}$ telles qu'il existe $\alpha > 0$ pour lequel on a :

$$\int_K f(x(t)/\alpha)\, dt < +\infty.$$

Comme dans le cas des suites, $L_f(K)$ est aussi l'ensemble des fonctions $x(t)$ pour lesquelles :

$$\sup\{|\int_K x(t)y(t)\,dt|; \int_K g(y(t))\,dt \leqslant 1\} < +\infty.$$

$L_f(K)$ muni de l'une des deux normes équivalentes :

$$\|x(t)\| = \inf\{\alpha ; \int_K f(x(t)/\alpha)\,dt \leqslant 1\}$$

ou :

$$|||x(t)||| = \sup\{b \left|\int_K x(t)y(t)\,dt\right|; \int_K g(y(t))\,dt \leqslant 1\}$$

est un espace de Banach appelé espace d'Orlicz de fonctions associé à la N-fonction f.

3. Cas général

Dans ce chapitre, E désigne l'espace $\mathbf{R}^n$ ou, plus généralement, un espace vectoriel topologique séparé localement convexe sur $\mathbf{R}$; dans ce dernier cas, le dual topologique E* de E sera muni de la topologie faible $\tau_s(E)$ donnée par E et E sera muni de la topologie faible $\tau_s(E^*)$ donnée par E*.

Il ne faudrait pas croire que l'on peut, comme dans le cas des fonctions convexes de $\mathbf{R}$ dans $\mathbf{R}$, conclure à la continuité des fonctions convexes de E dans $\overline{\mathbf{R}}$; on dispose, en fait, du résultat suivant :
Soit f une fonction convexe prenant une valeur finie en un point x de E ; s'il existe un voisinage de x sur lequel f est majorée par une constante finie, elle est continue au point x.

Ce résultat permet, dans le cas particulier où $E = \mathbf{R}^n$, d'établir que :
Toute fonction convexe propre sur $\mathbf{R}^n$ est continue sur l'intérieur de son domaine effectif ; en particulier, si f est à valeurs dans $\mathbf{R}$, alors dom $(f) = \mathbf{R}^n$ et f est continue sur $\mathbf{R}^n$.

Rappelons qu'une fonction f de E dans $\overline{\mathbf{R}}$ est dite semi-continue inférieurement si, pour tout réel a, l'ensemble des éléments x de E tels que $f(x) \leqslant a$ est fermé ; il est équivalent de dire que l'épigraphe de f est fermé, ou encore que f est enveloppe supérieure d'une famille de fonctions continues, c'est-à-dire que :

$$f = \sup_{\alpha \in a} f_\alpha,$$

où f_α est continue pour tout $\alpha \in a$.

L'importance des hyperplans d'appui dans l'étude des ensembles convexes nous amène à introduire pour une fonction convexe f de E dans $\overline{\mathbf{R}}$ la famille A_f des fonctions affines continues qui minorent f. Si on note $\Gamma(E)$ l'ensemble des fonctions h de E dans $\overline{\mathbf{R}}$ qui sont enveloppe supérieure d'une famille de fonctions affines continues, le théorème de séparation (cf. chap. 4 de la partie A ci-dessus – Ensembles convexes) permet de montrer que :
L'ensemble h est élément de $\Gamma(E)$ si et seulement si h est une fonction convexe propre semi-continue inférieurement ou vaut identiquement $-\infty$.

Pour une fonction f de E dans $\overline{\mathbf{R}}$, la plus grande fonction g de $\Gamma(E)$ qui minore f est aussi l'enveloppe supérieure des fonctions affines continues qui minorent f ; dans le cas où f est une fonction convexe et où A_f est non vide, g est aussi la plus grande fonction semi-continue inférieurement qui minore f ; si bien que, lorsque f est de plus semi-continue inférieurement, $f = g$.

Fonction conjuguée d'une fonction convexe

L'inégalité de Young du chapitre 2 s'écrit $xy - g(y) \leqslant f(x)$, ce que l'on peut interpréter en disant que la fonction affine $l(x) = xy - g(y)$ est une minorante de f ; la fonction $g(y)$ est choisie de telle sorte que $l(x)$ soit la plus grande minorante affine de f de coefficient directeur y.

Si, maintenant, f est une fonction convexe de E dans $\overline{\mathbf{R}}$, de la même façon, pour tout $x^* \in E^*$, introduisons la fonction affine $l(x) = x^*(x) - \alpha$ et cherchons à déterminer *si possible* α, de manière à obtenir la plus grande minorante affine de f de forme linéaire associée x^* ; cela nous conduit à introduire :

$$(10) \qquad f^*(x^*) = \sup_{x \in E} (x^*(x) - f(x)).$$

Si $f^*(x^*) \in \mathbf{R}$, alors la fonction $l(x) = x^*(x) - f^*(x^*)$ est effectivement la plus grande minorante affine continue de forme linéaire associée x^*.

La formule (10) généralise la transformation de Legendre (7) du chapitre 2 et permet de définir sur E* une fonction f^* convexe qui est dans la classe $\Gamma(E^*)$; la fonction f^* est, par définition, la fonction conjuguée de f.

Remarquons que si la fonction $f(x)$ est finie, de la formule (10) on tire l'inégalité de Young :

$$(11) \qquad f(x) + f^*(x^*) \geqslant x^*(x)$$

qui généralise l'inégalité (6) du chapitre 2.

Recommençons maintenant le procédé en posant, pour tout $x \in E$,

$$f^{**}(x) = \sup_{x^* \in E^*} (x^*(x) - f^*(x^*)) ;$$

f^{**} est alors le plus grand élément de $\Gamma(E)$ qui minore f; donc, si $f \in \Gamma(E)$, alors $f = f^{**}$.

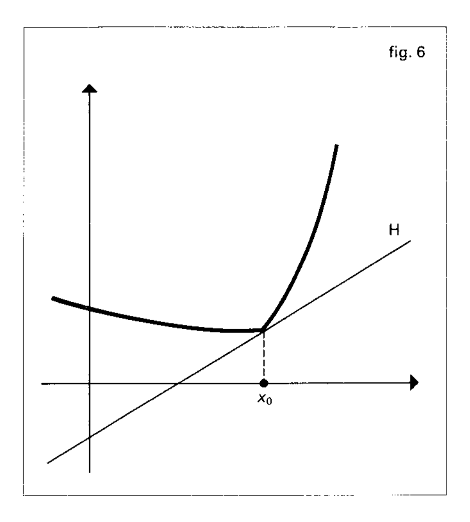
Sous-différentiel

Soit f une fonction convexe de E dans $\overline{\mathbf{R}}$. Supposons qu'il existe un élément l de A_f (c'est-à-dire une minorante affine continue de f) tel que $l(x_0) = f(x_0)$; on dit alors que f est *sous-différentiable* en x_0 ; l'application linéaire continue x^* associée à l'application affine l est un *sous-gradient* de f en x_0 ; l'ensemble des sous-gradients de f en x_0 est appelé le *sous-différentiel* en x_0 de f et est noté $\partial f(x_0)$.

Remarquons que, dans ces conditions, on a :

$$|f(x_0)| < +\infty ; \; l(z) = x^*(z - x_0) + f(x_0) ;$$

l'hyperplan H graphe dans $E \times \mathbf{R}$ de l est un hyperplan d'appui du convexe épi(f) (fig. 6). Comme on a vu, par ailleurs, que la plus grande minorante affine continue



de f ayant x^* pour application linéaire associée est donnée par $x^*(z) - f^*(x^*)$, on en conclut que $f^*(x^*) = x^*(x_0) - f(x_0)$; en fait, cette égalité est une conditon nécessaire et suffisante pour que $x^* \in \partial f(x_0)$.

Si f est dans $\Gamma(E)$ (cas où $f = f^{**}$), les propositions suivantes sont équivalentes :

(1) $\quad x^* \in \partial f(x)$,

(2) $\quad x \in \partial f^*(x^*)$,

(3) $\quad f(x) + f^*(x^*) = x^*(x)$,

(4) $\quad f(x) + f^*(x^*) - x^*(x) \leqslant 0$.

Le sous-différentiel en x_0 d'une fonction convexe f est un sous-ensemble convexe fermé de E*.

Le résultat suivant donne une condition intéressante de sous-différentiabilité d'une fonction convexe :

Si f est finie et continue en un point x_0, f est sous-différentiable en tout point intérieur de dom(f) et en particulier en x_0.

Dans le cas où f est Gâteaux-différentiable en x_0, c'est-à-dire s'il existe un élément $f'(x_0)$ de E* tel que, pour tout y de E, on ait :

$$\lim_{\substack{\lambda \to 0 \\ \lambda > 0}} \frac{f(x_0 + \lambda y) - f(x_0)}{\lambda} = f'(x_0)(y),$$

la fonction f est sous-différentiable en x_0 et $f'(x_0)$ est l'unique sous-gradient de f en x_0. Réciproquement si, en x_0, f est continue, sous-différentiable et ne possède qu'un seul sous-gradient x^*, alors f est Gâteaux-différentiable en x_0 et $f'(x_0) = x^*$.
Le sous-différentiel permet de remplacer la différentielle et d'exprimer notamment des conditions d'optimalité dans des problèmes de contrôle.

Donnons l'exemple de la fonction F, définie sur $L^2(\Omega)$, où Ω est un ouvert de $\mathbf{R}^n$ suffisamment régulier, par :

$$F(u) = \begin{cases} +\infty, \text{ si } u \notin W_0^{1,2}(\Omega) \\ \hbar \int_\Omega \sum_{i=1}^{n} \left| \frac{\partial u(x)}{\partial x_i} \right|^2 dx, \text{ si } u \in W_0^{1,2}(\Omega) \end{cases}$$

$w_0^{1,2}(\Omega)$ représente ici le sous-espace de l'espace de Sobolev $W^{1,2}(\Omega)$ constitué des u dont la restriction au bord de a est nulle. F est alors une fonction convexe semi-continue inférieurement, sous-différentiable en chaque point de $W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega)$; pour chaque point u de ce sous-espace, le sous-différentiel $\partial F(u)$ est constitué du seul élément :

$$u^* = -2 \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2},$$

c'est-à-dire que, pour $v \in L^2(\Omega)$, on a :

$$u^*(v) = -2 \int_\Omega \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2} v \, dx \cdot$$

Notons encore que, si f est une fonction convexe propre de $\Gamma(E)$ pour tous les x_1, x_2, x_1^*, x_2^* vérifiant $x_1^* \in \partial f(x_1)$ et $x_2^* \in \partial f(x_2)$, on a :

$$(x_1^* - x_2^*)(x_1 - x_2) \geqslant 0.$$

On dit que le sous-différentiel est un opérateur monotone ; il est même maximal monotone en ce sens que, pour tout couple(x, x^*) tel que $x^* \in \partial f(x)$, il existe un couple (y, y^*) tel que :

$$(y^* - x^*)(y - x) < 0.$$

La théorie des opérateurs maximaux monotones, qui généralise l'analyse convexe, est très utile pour l'étude des équations d'évolution non linéaires de type parabolique ou hyperbolique.

ROBERT ROLLAND

Bibliographie

I. Ekeland & R. Temam, *Analyse convexe et problèmes variationnels*, Dunod, Paris-Bruxelles-Montréal, 1974 / Y. Meyer, *Option analyse convexe de l'École polytechnique*, 1981 / T. Rockafellar, *Convex Analysis*, Princeton Univ. Press, 1974 / M. Willem, *Analyse convexe et optimisation*, C.I.A.C.O., Louvain-la-Neuve, 1987.

CORPS

La structure de corps n'est en fait qu'un cas particulier de la structure plus générale d'anneau (cf. ANNEAUX ET ALGÈBRES) ; en plus des axiomes généraux, on stipule que le groupe multiplicatif des éléments inversibles est le complémentaire de 0. Les corps sont donc les domaines dans lesquels les opérations habituelles du calcul sont valables, y compris la division par un élément non nul. La terminologie habituelle sous-entend la commutativité de la multiplication, mais il s'introduit de manière naturelle des corps où la multiplication n'est pas commutative (cf. *Quaternions*, in ANNEAUX ET ALGÈBRES, chap. 2 et *infra*, chap. 3). Du point de vue arithmétique, l'étude d'un corps commutatif se caractérise par l'absence d'idéaux non triviaux.

On se limitera ici à la théorie proprement algébrique des corps, mais on rencontre aussi des corps munis de structures additionnelles compatibles avec la structure de corps : les corps ordonnés, les corps topologiques et les corps valués (cf. algèbre TOPOLOGIQUE ; théorie des NOMBRES – Nombres p-adiques).

Un sous-ensemble K d'un corps L qui est un corps pour l'addition et la multiplication induites est appelé un *sous-corps* de L. Pour ne prendre que des exemples bien connus, les nombres rationnels forment un sous-corps **Q** du corps **R** des nombres réels, qui est lui-même un sous-corps du corps **C** des nombres complexes.

Si K apparaît comme *sous-corps* d'un corps L, on dit aussi que L est une *extension* de K. On peut alors considérer L comme un espace vectoriel à gauche sur K, l'opération externe n'étant autre que la multiplication à gauche des éléments de L par les éléments de K. Si cet espace vectoriel L est de dimension finie n sur K, on dit que L est une *extension finie* de K ; le nombre n s'appelle le *degré* de L sur K, et on le note [L : K]. Si M est une extension finie de L, c'est une extension finie de K et on a :

$$[\mathrm{M} : \mathrm{K}] = [\mathrm{M} : \mathrm{L}][\mathrm{L} : \mathrm{K}].$$

Un homomorphisme f d'un corps K dans un corps L est un homomorphisme d'anneau, c'est-à-dire qui respecte les deux lois additive et multiplicative, avec la condition importante $f(1) = 1$. Un tel homomorphisme est nécessairement *injectif* car tout $x \neq 0$ a un inverse x^{-1}, d'où $f(1) = f(xx^{-1}) = f(x)f(x)^{-1} = 1$, d'où $f(x) \neq 0$; ainsi f identifie K à un sous-corps $\mathrm{K}' = f(\mathrm{K})$ de L et réalise ainsi L comme une extension de K. Si cette injection est une bijection, f est un *isomorphisme*. Les isomorphismes d'un corps K sur lui-même, ou *automorphismes* du corps K, jouent un rôle particulièrement important dans l'étude de la structure du corps (cf. *Théorie de Galois*).



1. Exemples

Suffisamment « rigides » pour être maniés et étudiés précisément, les corps constituent à la fois un modèle et un outil qui interviennent dans de nombreux domaines des mathématiques et dans des questions, même relativement élémentaires, de géométrie algébrique, analytique ou projective ou de théorie des nombres. Voici quelques exemples.

Caractéristique d'un corps et corps finis

L'intersection d'une famille de sous-corps d'un corps K est encore un corps. Considérant en particulier la famille de *tous* les sous-corps de K, on obtient le plus petit sous-corps de K, appelé *sous-corps premier* K_0 de K. Notant $n.1$ la somme de n exemplaires de 1, pour tout entier naturel n, on définit la *caractéristique* (cf. ANNEAUX ET ALGÈBRES, chap. 3).

Si $n.1 \neq 0$ pour $n \neq 0$, on dit que K est de caractéristique nulle. Les $n.1$ et $-(n.1)$ pour $n \in \mathbf{N}$ forment donc un sous-anneau de K isomorphe à l'anneau **Z** des entiers relatifs et le corps K_0 est isomorphe au corps **Q** des nombres rationnels. Le corps K est donc une extension du corps des nombres rationnels.

Dans le cas contraire, la caractéristique de K est le plus petit entier strictement positif tel que $p.1 = 0$. C'est un nombre premier et le corps K_0 est alors isomorphe au corps fini $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$ des entiers relatifs

modulo p (cf. ANNEAUX ET ALGÈBRES, chap. 3). Ainsi, tout corps de caractéristique p est une extension du corps $\mathbf{F}_p$ et deux corps de caractéristiques différentes ne peuvent être extension l'un de l'autre.

Soit K un corps fini. Un théorème dû à J. H. M. Wedderburn affirme qu'un tel corps est nécessairement commutatif. La caractéristique de K est nécessairement un nombre premier p et K est une extension finie du corps premier $\mathbf{F}_p$. Si $n = [K : \mathbf{F}_p]$, alors K est isomorphe à $(\mathbf{F}_p)^n$ comme espace vectoriel sur $\mathbf{F}_p$ et il a donc p^n éléments. On verra ci-dessous que pour tout entier de la forme p^n, avec n premier, il existe un corps (unique à un isomorphisme près) possédant p^n éléments ; on le note $\mathbf{F}_{p^n}$.

Corps de nombres

Le corps **C** des nombres complexes est un exemple bien classique de corps. Les sous-corps de **C** forment une vaste famille à laquelle appartiennent le corps **Q** des nombres rationnels (qui est le plus petit) et le corps **R** des nombres réels. Les corps de nombres algébriques présentent un intérêt tout particulier. Dedekind en donne la description suivante : Soit x un nombre complexe algébrique, c'est-à-dire une racine d'une équation $P(X) = 0$, où $P(X)$ est un polynôme à coefficients entiers, de degré n irréductible sur le corps **Q** ; alors l'ensemble $\mathbf{Q}(x)$ des nombres complexes de la forme :

$$a_0 + a_1 x + \ldots + a_{n-1} x^{n-1},$$

où les a_i sont des nombres rationnels quelconques, est un corps. De la définition, il résulte que $\mathbf{Q}(x)$ est un espace vectoriel sur **Q** de dimension finie n. Inversement, on peut montrer que toute extension finie de **Q** est isomorphe à une extension de la forme précitée $\mathbf{Q}(x)$. Si bien que l'on peut définir abstraitement les corps de nombres algébriques comme des extensions finies de **Q**. Ainsi, si, dans l'anneau $\mathbf{Q}[X]$ des polynômes à coefficients rationnels, on identifie deux polynômes $R(X)$ et $R'(X)$ dont la différence est un multiple d'un polynôme $P(X)$, à coefficients entiers, de degré n, irréductible sur **Q**, on obtient, sur l'ensemble quotient $\mathbf{Q}[X]/P(X)$ muni de l'addition et de la multiplication induites par celles des polynômes, une structure de corps qui en fait une extension finie de degré n de Q. En choisissant une racine x de l'équation $P(X) = 0$ dans le corps des nombres complexes, on peut expliciter un isomorphisme de $\mathbf{Q}[X]/(P(X))$ sur $\mathbf{Q}(x)$ défini précédemment : à un polynôme $R(X)$ on associe sa valeur $R(x)$ en x et, comme deux polynômes congrus *modulo* $P(X)$ ont même valeur en x, cela définit un homomorphisme :

$$J : \mathbf{Q}[X]/(P(X)) \to \mathbf{Q}(x),$$

qui est l'isomorphisme annoncé. La dernière définition des corps de nombres algébriques, qui est, au langage près, celle de Kronecker, est ainsi reliée à celle de Dedekind.

Corps de restes

Le procédé de Kronecker pour définir les corps de nombres algébriques peut être présenté dans un contexte plus général. Un idéal m d'un anneau commutatif unitaire A est appelé *idéal maximal* s'il n'est contenu strictement dans aucun autre idéal que A lui-même. L'anneau quotient A/m ne possède alors aucun idéal autre que 0 et A/m, car de tels idéaux sont en correspondance biunivoque avec les idéaux de A qui contiennent m. Tout élément non nul x de A/m engendre donc A/m tout entier ; il en résulte qu'il existe un élément x^{-1} de A/m tel que $xx^{-1} = 1$ et que l'anneau unitaire

A/m est en fait un corps, le *corps des restes de* A *modulo m*.

Nous avons déjà appliqué ce résultat au plus simple de tous les anneaux unitaires : l'anneau **Z** des entiers relatifs. Les idéaux maximaux de **Z** sont les idéaux p**Z** engendrés par un nombre premier p et le corps des restes $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$ possède p éléments. Nous avons ici un premier exemple de corps à un nombre fini d'éléments, ou *corps finis*. Nous reviendrons sur ces corps (parfois appelés champs de Galois dans la vieille littérature), dont l'importance est essentielle en théorie des nombres.

Dans l'anneau K[X] des polynômes à une variable sur un corps commutatif K, un idéal est maximal si, et seulement si, il est engendré par un polynôme irréductible non constant P(X). Les classes de polynômes modulo P(X) forment donc un corps K[X]/(P(X)). C'est ainsi que le corps des nombres complexes peut être défini, avec Cauchy, comme le corps de restes $\mathbf{R}[X]/(X^2 + 1)$. Si $K = \mathbf{Q}$, on retrouve les corps de nombres algébriques de Kronecker.

Corps de fractions

Tout corps possède la propriété que le produit de deux éléments non nuls est lui-même non nul ; il en est de même de tout sous-anneau (c'est-à-dire de tout sous-ensemble du corps qui, pour l'addition et la multiplication induites, est un anneau). Un anneau possédant une telle propriété est appelé un anneau intègre. La construction des nombres rationnels à partir des entiers relatifs suggère un moyen de considérer tout anneau commutatif intègre A comme un sous-anneau d'un corps K. On considère d'abord, sur $A \times (A - \{0\})$, les lois de composition :

$$(a,b) + (a',b') = (ab' + ba', bb'),$$
$$(a,b)(a',b') = (aa', bb'),$$

puis on vérifie que la relation d'équivalence $\mathcal{R}$, définie sur $A \times (A\text{-}\{0\})$ par $(a,b)\mathcal{R}(a',b')$, lorsque $ab' = ba'$, est compatible avec ces lois de composition et que le quotient K est un corps : l'unité est la classe de (a,a), et l'inverse de la classe de (a,b) existe dès que a n'est pas nul et n'est autre que la classe de (b,a).

Il est d'usage de noter a/b la classe d'un couple (a,b). L'anneau A s'identifie alors au sous-anneau de K formé des classes du type $a/1$. Il est à remarquer que cette construction est « universelle » : chaque fois que A sera obtenu comme sous-anneau d'un corps K′, le corps K′ pourra être considéré comme extension du corps K. On dit que K est *le corps des fractions* de A.

Nous allons appliquer ce qui précède à deux importants cas particuliers. Si K est un corps commutatif, l'anneau K[X] des polynômes :

$$P(X) = a_0 + a_1X + ... + a_nX^n,$$

à coefficients dans K est intègre. Il est alors possible de former le corps des fractions de K[X], noté K(X), et dont les éléments P(X)/Q(X), où P(X) et Q(X) sont deux polynômes, sont appelés *fractions rationnelles* sur K. Il est facile de généraliser cela au cas de plusieurs variables : on obtient alors le corps $K(X_1, X_2, ..., X_n)$ des fractions rationnelles à n variables indéterminées comme corps des fractions de l'anneau intègre $K[X_1, X_2, ..., X_n]$ des polynômes à n variables.

De même, l'anneau des séries formelles entières :

$$S(X) = a_0 + a_1X + ... + a_nX^n + ...,$$

à coefficients dans K (cf. ANNEAUX ET ALGÈBRES), que l'on note habituellement K[[X]], est intègre. Il est donc de nouveau possible de former le corps K((X)) des fractions de K[[X]]. Si l'on remarque que

les séries formelles entières dont le terme constant a_0 n'est pas nul sont inversibles dans K[[X]], on voit que toute série formelle entière non nulle peut se mettre, d'une façon unique, sous la forme X^n S(X), où S(X) est une série formelle entière inversible. Pour construire le corps K((X)), il suffit donc de savoir effectuer la division par les monômes X^n. Il en résulte immédiatement que le corps K((X)) peut être décrit comme l'ensemble des séries formelles :

$$S(X) = \sum_{n \gg -\infty} a_n X^n$$

(où la notation $n \gg -\infty$ signifie que, pour n inférieur à un certain entier relatif n(S), on a $a_n = 0$) muni des lois d'addition et de multiplication qui prolongent celles des séries formelles entières :

$$\left(\sum_{m \gg -\infty} a_m X^m\right) + \left(\sum_{n \gg -\infty} b_n X^n\right) = \sum (a_r + b_r) X^r,$$

$$\left(\sum_{m \gg -\infty} a_m X^m\right) \times \left(\sum_{n \gg -\infty} b_n X^n\right) = \sum c_r X^r,$$

$$\text{où } c_r = \sum_{m+n=r} a_m b_n .$$

Corps de fonctions algébriques

La géométrie algébrique fournit de nombreux exemples de corps. Nous nous limiterons ici à des indications élémentaires.

Une *sous-variété algébrique* affine de l'espace vectoriel $\mathbf{C}^n$ des suites $(x_1, x_2, ..., x_n)$ de n nombres complexes est définie comme l'ensemble V des points $(a_1, a_2, ..., a_n)$ de $\mathbf{C}^n$ qui vérifient un certain nombre d'équations :

$$\begin{cases} P_1(a_1, a_2, ..., a_n) = 0, \\ \dots\dots\dots\dots \\ \dots\dots\dots\dots \\ P_r(a_1, a_2, ..., a_n) = 0, \end{cases}$$

où les $P_i(X_1, X_2, ..., X_n)$ sont des polynômes à coefficients complexes. Un polynôme $S(X_1,X_2,...,X_n)$ à coefficients complexes définit une fonction à valeurs complexes sur $\mathbf{C}^n$ et, par restriction, sur V. On considérera comme équivalents deux polynômes qui prennent la même valeur aux points de V. L'ensemble des classes d'équivalence, noté C[V], est muni d'une addition et d'une multiplication, déduites de celles des polynômes, qui en font un anneau, et ses éléments peuvent être considérés comme des fonctions sur V à valeurs complexes.

Si V n'est pas la réunion de deux sous-variétés V′ et V″ distinctes d'elle-même, on dit qu'elle est *irréductible.* Dans ce cas, l'anneau C[V] est intègre, et le corps des fractions, qui est noté C(V), peut être vu comme un corps de fonctions sur V, à valeurs dans la droite projective complexe (ensemble obtenu par l'adjonction à **C** d'un point « à l'infini »). Les éléments de C(V) sont appelés des *fonctions algébriques* sur V.

2. Théorie élémentaire des corps commutatifs

Adjonction, extensions simples

Soit L un corps et K un sous-corps de L. Pour tout sous-ensemble S de L, l'intersection des sous-corps de L qui contiennent K et S est un sous-corps de L, que l'on appelle le sous-corps obtenu par *adjonction* de S à K, et que l'on note K(S). Si K(S) = L, on dit que S est un *système de générateurs* de L sur K. Un cas particulier important est celui où S est réduit à un seul élément x : l'extension obtenue est notée K(x), et on dit que c'est une *extension simple* de K. En effet, toute extension L d'un corps K peut être obtenue par adjonc-

tions « répétées » d'un élément (lorsque L possède un système S fini ou dénombrable de générateurs, l'expression « répétées » a le sens classique ; pour la définir dans le cas général, il faut faire une récurrence transfinie après avoir muni S d'un bon ordre, ce qu'autorise l'axiome de Zermelo).

Soit L une extension simple d'un corps K, et soit x un générateur, c'est-à-dire que $L = K(x)$. On peut faire un raisonnement tout à fait parallèle à celui qui a été fait à propos de la caractéristique. En effet, deux cas se présentent :

– Les monômes x^n sont linéairement indépendants sur K, c'est-à-dire qu'une relation telle que :

$$a_0 + a_1x + a_2x^2 + \dots + a_nx^n = 0,$$

avec les a_i dans K, n'est possible que si $a_0 = a_1 = a_2 = \dots = a_n = 0$. Le corps $K(x)$ est alors isomorphe au corps $K(X)$ des fractions rationnelles sur K. On dit que x est *transcendant* et que $K(x)$ est une *extension transcendante simple* de K. Évidemment, tout élément y de $K(x)$ qui n'appartient pas à K est transcendant sur K et on a $K(y) = K(x)$. Les exemples classiques de nombres transcendants sur Q sont ceux de $e = 2{,}718\,28\dots$, base des logarithmes népériens, et $\pi = 3{,}141\,59\dots$, rapport de la circonférence d'un arc à son diamètre (cf. nombres TRANSCENDANTS).

– Il existe un polynôme non constant, que l'on peut supposer irréductible, $P(X)$, à coefficients dans K tel que $P(x) = 0$. Le corps $K(x)$ est alors isomorphe au corps de restes $K[X]/(P(X))$. On dit que x est *algébrique* sur K et que $K(x)$ est une *extension algébrique simple* de K. Si le polynôme $P(X)$, que l'on appelle *polynôme minimal* de x, est de degré n, alors $(1, x, x^2, \dots, x^{n-1})$ est une base de $K(x)$ sur K et on a donc $[K(x) : K] = n$.

Extensions algébriques, bases de transcendance

On peut généraliser ce qui vient d'être dit au précédent paragraphe. Un élément x d'une extension L d'un corps K est *algébrique* s'il vérifie une équation algébrique à coefficients dans K :

$$x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 = 0,$$

ou, en d'autres termes, si l'extension simple $K(x)$ de K est algébrique. Si tous les éléments de L sont algébriques sur K, on dit que L est une *extension algébrique* de K. Ainsi en est-il d'une extension algébrique simple. On peut montrer facilement que toute extension algébrique engendrée par un nombre fini d'éléments est finie. La réciproque étant claire, il n'y a pas lieu de distinguer les *extensions finies* de celles que l'on appelle parfois *extensions algébriques finies*.

Revenons maintenant au cas général d'une extension quelconque L d'un corps K. Un sous-ensemble S de L est *algébriquement indépendant sur* K, par définition, si, pour tout sous-ensemble fini $(s_1, s_2, \dots, s_n)$ de S, il n'existe aucun polynôme à coefficients dans K non nul $P(X_1, X_2, \dots, X_n)$, tel que $P(s_1, s_2, \dots, s_n) = 0$. Le corps $K(s_1, s_2, \dots, s_n)$ engendré sur K par les s_i est alors isomorphe au corps des fractions rationnelles à n variables sur K, $K(X_1, X_2, \dots, X_n)$. Une *base de transcendance* de L sur K est un sous-ensemble T de L, algébriquement indépendant sur K, et tel que L soit une extension algébrique de $K(T)$. On démontre qu'il existe toujours de telles bases de transcendance et que deux bases quelconques ont le même nombre d'éléments ; on appelle ce nombre le *degré de transcendance de* L *sur* K. Une extension est algébrique si, et seulement si, son degré de transcendance est nul.

Nous avons parlé plus haut du corps C(V) des fonctions algébriques rationnelles sur une sous-variété algébrique V de $\mathbf{C}^n$. Le degré de transcendance de C(V) sur **C** est égal à la dimension de V considérée comme variété analytique complexe (dimension elle-même égale à la moitié de la dimension de V considérée comme variété différentiable, mais cela est une autre histoire !). En particulier, si V est une courbe, le degré de transcendance est 1. On appelle parfois une extension de degré de transcendance 1 de **C**, et plus généralement d'un corps K, un *corps de fonctions* sur K.

Corps algébriquement clos, clôture algébrique, corps de rupture

Une équation algébrique à coefficients dans un corps K n'admet pas nécessairement de racine dans K. Ainsi, l'équation à coefficients réels $X^2 + 1 = 0$ n'a pas de racine réelle. De même, dans **Z**/2**Z**, le polynôme $X^2 + X + 1$ prend la valeur 1 sur les deux éléments 0 et 1 et n'a donc aucun zéro. Si un corps K est tel que tout polynôme à coefficients dans K admette une racine dans K, on dit qu'il est *algébriquement clos*. Un tel corps ne saurait avoir d'extension algébrique propre ; inversement, un corps qui n'admet pas d'extension algébrique propre est algébriquement clos. Dans un corps algébriquement clos, un polynôme non constant se décompose en un produit de facteurs (irréductibles) du premier degré. Un théorème, démontré par Gauss (cf. nombres COMPLEXES), montre que le corps **C** des nombres complexes est algébriquement clos. Le mathématicien allemand Steinitz, qui a exposé vers 1910 une nouvelle théorie des corps commutatifs, a démontré que tout corps K admettait au moins une extension $\overline{K}$ qui soit algébrique sur K et algébriquement close. On appelle une telle extension une *clôture algébrique de* K. Si $\overline{K}_1$ et $\overline{K}_2$ sont deux clôtures algébriques de K, il existe un isomorphisme de $\overline{K}_1$ sur $\overline{K}_2$ qui laisse fixe chaque élément du sous-corps K, si bien que, dans la pratique, on ne les distingue pas. Le corps **C** est algébrique sur **R** et est donc une clôture algébrique de **R** ; mais ce n'est pas une clôture algébrique de **Q**, car des éléments tels que π et e sont transcendants sur **Q**. Le sous-corps $\overline{\mathbf{Q}}$ de **C**, formé par l'ensemble des nombres complexes algébriques sur **Q**, est algébriquement clos, et, comme, par définition, il est algébrique sur **Q**, il en est une clôture algébrique.

Pour le corps fini $\mathbf{F}_p$, on a :

$$\overline{\mathbf{F}}_p = \bigcup_n \mathbf{F}_{p^n},$$

où : $$\mathbf{F}_{p^n} = \{x \in \overline{\mathbf{F}}_p \mid x^{p^n} - x = 0\}$$

(cf. *Théorie de Galois*).

On connaît aussi la structure de la clôture algébrique du corps K((X)) des séries formelles à coefficients dans un corps K algébriquement clos de caractéristique nulle. Elle s'obtient par une méthode analogue à $\overline{\mathrm{F}}_{\mathrm{p}}$. On a :

$$\overline{\mathrm{K}((\mathrm{X}))} = \bigcup_n \mathrm{K}((\mathrm{X}^{1/n})),$$

où $\mathrm{K}((\mathrm{X}^{1/n}))$ est l'extension de degré n de K((X)) obtenue en adjoignant les racines de l'équation algébrique $\mathrm{T}^n - \mathrm{X} = 0$. Ces corps sont appelés corps de Puiseux.

Étant donné un corps K, $\overline{K}$ une clôture algébrique de K, les racines dans $\overline{K}$ d'un polynôme P(X) à coefficients dans K engendrent sur K un corps K_P. Tout corps dans lequel P(X) se décompose en facteurs du premier degré peut être considéré comme une extension de K_P : on dit que K_P est un *corps de rupture* pour le polynôme P(X). Comme le corps K_P est

engendré par un nombre fini d'éléments algébriques sur K, c'est une extension finie de K.

Automorphismes, extensions normales, groupes de Galois

Un *K-automorphisme* d'une extension L d'un corps K est un automorphisme σ du corps L tel que, pour tout x dans K, on ait $x^\sigma = x$ (nous utilisons la notation exponentielle, et le composé $\sigma\tau$ de deux automorphismes σ et τ est défini par $y^{\sigma\tau} = (y^\sigma)^\tau$). Ainsi, tout automorphisme d'un corps K est un K_0-automorphisme, K_0 étant le sous-corps premier de K. On notera G(L/K) le groupe des K-automorphismes d'une extension L d'un corps K. Pour tout sous-groupe H de G(L/K), on peut considérer l'ensemble L^H des éléments de L laissés fixes par tout automorphisme appartenant à H. Il est immédiat que L^H est un corps qui contient K. On l'appelle le *corps des invariants* de H. Voici deux exemples de cette situation :

- La conjugaison complexe σ qui, à tout nombre $x = a + ib$ de **C**, associe le nombre $\bar{x} = a - ib$, est un automorphisme du corps **C**. Le groupe H d'automorphismes de **C** formé par σ et l'identité admet **R** pour corps des invariants.
- Dans un corps K de caractéristique p non nulle, l'application φ de K dans K qui, à tout élément x, associe sa puissance p-ième x^p est un endomorphisme du corps, que l'on appelle *endomorphisme de Frobenius*. Le seul point non trivial à vérifier est que :

$$(x - y)^p = x^p - y^p\,;$$

mais cela résulte du fait que dans la formule du binôme, les coefficients non extrêmes sont divisibles par p puisque p est premier. Lorsque cet endomorphisme est un automorphisme, ce qui est le cas pour les corps finis, on dit que le corps K est *parfait*. On peut alors chercher quels sont les invariants dans K pour le groupe d'automorphismes H engendré par φ : ce sont les racines de l'équation : $X^p - X = 0$; il y en a donc au plus p, et K^H n'est autre que le sous-corps premier K_0, qui est isomorphe au corps à p éléments $\mathbf{Z}/p\mathbf{Z}$.

Il est clair que les K-automorphismes d'un corps L respectent le caractère de transcendance ou d'algébricité sur K des éléments de L. Le dernier point peut être précisé : si x est racine d'une équation algébrique :

$$a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0$$

irréductible sur K, on a, en posant $y = x^\sigma$, $\sigma \in G(L/K)$:

$$\begin{aligned}&a_n y^n + a_{n-1}y^{n-1} + \ldots + a_1 y + a_0\\ &= (a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0)^\sigma = 0^\sigma = 0,\end{aligned}$$

si bien que x et son transformé x^σ ont même polynôme minimal (deux éléments algébriques de L qui sont dans ce cas sont dits *conjugués* sur K).

Une *extension algébrique normale* L d'un corps K est, par définition, une extension algébrique dans laquelle le polynôme minimal de tout élément x se décompose en facteurs du premier degré (en d'autres termes, L contient tous les conjugués de x dans une clôture algébrique L de L). Il est évident qu'un corps de rupture sur un corps K d'un polynôme P(X) à coefficients dans K est une extension algébrique normale de K. Le groupe G(L/K) d'une extension algébrique normale L d'un corps K, que l'on appelle alors le *groupe de Galois* de l'extension, opère transitivement dans toute classe d'éléments conjugués, c'est-à-dire que, si x et y sont deux éléments conjugués, il existe $\sigma \in G(L/K)$ tel que $y = x^\sigma$. Lorsque L est

le corps de rupture sur K d'un polynôme P(X) à coefficients dans K, le groupe G(L/K) est parfois nommé *groupe de l'équation* P(X) = 0.

À titre d'exemple, remarquons que le corps $\mathbf{Q}(\sqrt[3]{2})$ n'est pas une extension normale de **Q** : en effet, $\alpha = \sqrt[3]{2}$ a deux conjugés αj et $\alpha j'$, où j et j' sont les racines cubiques non réelles de l'unité. Le corps $Q(\alpha, j)$ obtenu par adjonction de j à $Q(\alpha)$ est un corps de rupture de $P(X) = X^3 - 2$, le polynôme minimal de α, et c'est aussi la plus petite extension de $\mathbf{Q}(\alpha)$ normale sur **Q**.

Théorie de Galois

Jusqu'à Abel et Galois, le problème central posé par les équations algébriques était celui de leur solution par radicaux, c'est-à-dire l'expression des racines au moyen d'opérations rationnelles et d'extractions de racines (cf. ÉQUATIONS ALGÉBRIQUES). Les Grecs connaissaient déjà des cas particuliers de la formule $x = (-b \pm \sqrt{b^2 - 4ac})/(2a)$ pour la solution de l'équation du second degré $ax^2 + bx + c = 0$, et de semblables formules avaient été trouvées pour les équations du troisième et quatrième degré par J. Cardano, N. Tartaglia et L. Ferrari. Les échecs répétés pour parvenir à une solution dans le cas de l'équation du cinquième degré amenèrent Lagrange (1770) à examiner avec plus de profondeur ce qui permettait d'arriver au but jusqu'au degré 4. C'est ainsi qu'il fut conduit à mettre en évidence certaines fonctions rationnelles des racines qui restaient invariantes par certaines substitutions effectuées sur celles-ci : les résolvantes qui portent son nom. Mais ce ne sera qu'avec Abel, et surtout Galois (1832), que l'on considérera les fonctions rationnelles des racines d'une équation polynomiale P(X) = 0 et les opérations d'addition et de multiplication sur celles-ci. De plus, Galois sut dégager un sous-groupe significatif du groupe de toutes les permutations des racines de l'équation, le groupe de l'équation, et lire sur ce groupe, entre autres choses, la possibilité ou l'impossibilité de résoudre l'équation par radicaux. Son résultat avait pour corollaire le théorème pressenti par Ruffini puis démontré par Abel : l'équation générale du cinquième degré n'est pas résoluble par radicaux. C'est Dedekind, qui a, comme nous l'avons déjà dit, introduit le terme de corps (pour les corps de nombres algébriques), et c'est lui encore qui a présenté le groupe d'une équation comme un groupe d'automorphismes de corps.

Exposés en un langage moderne, les résultats de Galois concernent une extension finie, normale et *séparable* (une extension algébrique est dite séparable si le polynôme minimal d'un élément quelconque n'a pas de racines multiples : toutes les extensions algébriques d'un corps de caractéristique 0 ou d'un corps parfait sont séparables). Dans la suite, nous dirons qu'une telle extension est *galoisienne*.

Le premier théorème précise la définition des extensions galoisiennes ; une extension finie L d'un corps K, de degré n, est galoisienne si, et seulement si, le corps L^G des invariants du groupe de Galois G= G(L/K) est réduit à K ; dans ce cas, le groupe de Galois G est d'ordre n.

Le deuxième théorème a trait à ce qu'on appelle la *correspondance de Galois*. Une extension galoisienne L d'un corps K étant donnée, l'application de l'ensemble des sous-groupes du groupe de Galois G(L/K) dans l'ensemble des sous-corps de L qui contiennent K, qui, à un sous-groupe H de G(L/K), associe le corps des invariants L^H, est bijective. L'application inverse peut être décrite ainsi : si M est un sous-corps

de L qui contient K, l'extension L du corps M est encore normale et séparable, donc galoisienne ; son groupe de Galois G(L/M) est un sous-groupe du groupe G(L/K), et l'application $M \mapsto G(L/M)$ est inverse de l'application $H \mapsto L^H$. De plus, un sous-corps M de L qui contient K est une extension galoisienne de K si, et seulement si, le groupe de Galois G(L/M) est un sous-groupe normal de G(L/K) et, dans ce cas, le groupe de Galois G(M/K) s'identifie au groupe quotient G(L/K)/G(L/M).

Il est maintenant possible de donner un sens précis à la « résolubilité par radicaux ». Une équation algébrique $P(X) = 0$ à coefficients dans un corps K de caractéristique 0 est *résoluble par radicaux*, par définition, si le corps de rupture K_p de P(X) sur K peut être « plongé » comme sous-corps dans un corps L tel qu'il existe une suite de sous-corps de L, $L_0 = K$, L_1, ..., L_{k-1}, $L_k = L$, avec $L_{i+1} = L_i(x_i)$, x_i étant racine d'une équation $x^{ni} - a_i = 0$ avec $a_i \in L_i$. En traduisant cette définition au moyen du dictionnaire que fournit la correspondance de Galois, on obtient assez facilement le critère : la résolubilité par radicaux de l'équation $P(X) = 0$ équivaut à la résolubilité du groupe de l'équation $G = G(K_p/K)$. Rappelons qu'un groupe G est un groupe *résoluble* s'il possède une suite de composition :

$$\{1\} = G_0 \subset G_1 \subset G_2 \subset ... \subset G_{n-1} \subset G_n = G,$$

telle que les quotients G_i/G_{i-1} soient des groupes commutatifs.

L'équation générale du n-ième degré : $X^n + t_{n-1} X^{n-1} + ... + t_1 X + t_0 = 0$, à coefficients algébriquement indépendants dans le corps $Q(t_0, t_1, ..., t_{n-1})$, a pour groupe le groupe symétrique S_n des permutations de n éléments. Comme on sait que ce groupe n'est pas résoluble lorsque $n \geqslant 5$, il est inutile d'espérer une formule de résolution par radicaux des équations de degré supérieur ou égal à 5. La théorie de Galois a permis de ramener le théorème de Ruffini-Abel à un théorème de théorie des groupes.

Signalons pour terminer qu'une extension L d'un corps K est dite *cyclique* (resp. *abélienne*, *résoluble*) si elle est galoisienne et si son groupe de Galois est cyclique (resp. commutatif, résoluble). L'étude des extensions abéliennes des corps de nombres algébriques constitue l'objet de la théorie du *corps de classes*, dont l'initiateur fut D. Hilbert (1900) et dont les principaux résultats furent démontrés par T. Takagi et E. Artin, 1920-1930 (cf. théorie des NOMBRES – Nombres algébriques).

Exemple de détermination d'un groupe de Galois

Nous avons déjà considéré le corps de rupture L de $P(X) = X^3 - 2$ sur le corps **Q** des nombres rationnels. Comme d'habitude, L sera vu comme un sous-corps de **C**. Soit α la racine réelle de $P(X) = 0$. Les conjugués de α sont αj et αj^2, où

$$j = \frac{-1 + i\sqrt{3}}{2} \text{ et } j^2 = \bar{j} = \frac{-1 - i\sqrt{3}}{2}$$

sont les racines cubiques non réelles de l'unité. Nous avons donc $L = \mathbf{Q}(\alpha, j)$. Comme $[L : \mathbf{Q}] = 6$, l'ordre du groupe de Galois G(L/**Q**) est 6. Il reste donc, pour déterminer complètement ce groupe, à trouver six automorphismes distincts de L. Un automorphisme de L est connu dès que l'on connaît son action sur α et j. La conjugaison complexe σ échange j et $\bar{j}$ et laisse fixe α ; l'automorphisme τ échange j et $\bar{j}$ d'une part et α et αj de l'autre. Nous pouvons dresser un tableau donnant les

images de α et j par différents automorphismes :

	ε	σ	τ	$\sigma\tau$	$\tau\sigma$	$\sigma\tau\sigma$
α	α	α	αj	$\alpha \bar{j}$	$\alpha \bar{j}$	$\alpha \bar{j}$
j	j	$\bar{j}$	$\bar{j}$	j	j	$\bar{j}$

Dans cet exemple, le groupe de Galois est le groupe de toutes les permutations des racines, mais, en général, c'est seulement un sous-groupe de ce corps, car toute permutation des racines ne se prolonge pas nécessairement en un automorphisme des corps.

Corps finis

Une application intéressante de la théorie de Galois est l'étude et la classification des corps finis. Soit donc F un groupe fini possédant $q = p^n$ éléments (cf. chap. 1). Le groupe multiplicatif des éléments non nuls de F est d'ordre $q-1$, donc tout élément de ce groupe vérifie $x^{q-1} - 1 = 0$ et, par suite, tout élément de F vérifie :

$$P_n(x) = x^{p^n} - x = 0.$$

Comme il est clair que les racines de $P_n(X)$ dans une clôture algébrique $\overline{\mathbf{F}}_p$ de $\mathbf{F}_p$ forment un corps, le corps à $q = p^n$ éléments F est un corps de rupture sur $\mathbf{F}_p$ pour le polynôme $P_n(X)$. Ce qui démontre l'existence et l'unicité, à un isomorphisme près, de corps finis à p^n éléments.

Nous pouvons interpréter ce qui précède en termes de théorie de Galois. Soit $\overline{\mathbf{F}}_p$ une clôture algébrique du corps premier $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$, et soit φ l'automorphisme de Frobenius qui associe à tout élément x de $\overline{\mathbf{F}}_p$ sa puissance p-ième x^p. Si G_n désigne le sous-groupe de $G(\overline{\mathbf{F}}_p/\mathbf{F}_p)$ engendré par φ^n, le corps des invariants de G_n n'est autre qu'un corps de rupture pour $P_n(X)$ que nous noterons $\mathbf{F}_{p^n}$, et qui a p^n éléments. L'automorphisme de Frobenius φ, considéré comme un automorphisme de $\mathbf{F}_{p^n}$ sur $\mathbf{F}_p$, est d'ordre n et il engendre donc le groupe de Galois $G(\mathbf{F}_{p^n}/\mathbf{F}_p)$. Soit plus généralement m et n deux entiers $\geqslant 1$, le corps $\mathbf{F}_{p^m}$ est extension du corps $\mathbf{F}_{p^n}$ si, et seulement si, n divise m ; cette extension est alors cyclique de degré m/n et son groupe de Galois est engendré par φ^n.

3. Corps non commutatifs

On a examiné jusqu'à présent des corps qui étaient commutatifs, mais l'étude des corps non commutatifs n'est pas d'un moindre intérêt.

Si K est un corps non commutatif, l'ensemble Z des éléments de K qui permutent avec tout élément x, c'est-à-dire tels que $xz = zx$, est visiblement un corps commutatif que l'on appelle le *centre* de K. Nous avons déjà signalé l'exemple des quaternions H dont le centre n'est autre que le corps R des nombres réels. Voici un autre exemple dû à Hilbert :

Soit $\mathbf{F}_q$ un corps fini à $q = p^r$ éléments ($r \geqslant 2$), si on munit l'ensemble des séries formelles $\sum_{n \gg +\infty} a_n T^n$ sur $\mathbf{F}_q$ de l'addition habituelle et d'une multiplication déduite par distributivité et associativité de la règle élémentaire $Ta = a^p T$, on obtient un corps non commutatif $\mathbf{F}'_q((T))$. Il est facile de voir que le centre de ce corps est formé des séries formelles constantes a_0 où $a_0 \in \mathbf{F}_p$ et qu'il est donc isomorphe à $\mathbf{F}_p$. Les séries formelles à coefficients dans $\mathbf{F}_p$ forment un sous-corps commutatif de $\mathbf{F}'_q((T))$.

On peut développer au sujet des corps non commutatifs des considérations tout à fait analogues à celles qui ont été faites dans le cas des corps commutatifs. En particulier, on sait définir la *caractéristique*

d'un corps non commutatif, et cette caractéristique n'est autre que celle du centre. De même, si L est un corps non commutatif et K un sous-corps de L, la notion d'*adjonction* à K d'un sous-ensemble S de L garde tout son sens. Il est à remarquer que si K est commutatif et x un élément de L qui permute avec tout élément de K, le sous-corps K(x) de L obtenu par adjonction de x à K est encore commutatif, ce qui permet de démontrer l'existence de *sous-corps commutatifs maximaux* dans L (c'est-à-dire de sous-corps commutatifs qui ne sont contenus strictement dans aucun autre sous-corps commutatif). L'étude des automorphismes des extensions commutatives finies d'un corps commutatif conduit à ce qu'on appelle la théorie de Galois. Mais il existe une théorie de Galois non commutative due à E. Noether et T. Skolem (1928), dont on donne ci-dessous quelques résultats. Si K est un corps non commutatif de centre Z, il est facile de mettre en évidence des automorphismes de K qui laissent Z fixe : pour tout élément non nul x de K, l'application $y \mapsto xyx^{-1}$ de K dans K est un automorphisme σ^x de K. Les automorphismes de la forme σ^x sont appelés les *automorphismes intérieurs* de K. Un théorème de Skolem-Noether assure que, si K est un corps non commutatif de degré fini sur son centre Z, il n'existe pas d'autres automorphismes de K laissant Z fixe que les automorphismes intérieurs. Un autre théorème permet de préciser la structure des corps non commutatifs K de degré fini sur son centre Z. Si L est un sous-corps de K qui contient Z, l'ensemble L′ des éléments y de K tels que $xy = yx$ pour tout x dans L est un sous-corps de K qui contient Z et que l'on appelle le *commutant* de L. On voit facilement que, si on répète l'opération, on a $(L')' = L$. De plus, on a l'égalité $[K : Z] = [L : Z]\,[L' : Z]$. Il résulte immédiatement des définitions qu'un sous-corps L est commutatif si, et seulement si, $L \subset L'$ et que les sous-corps commutatifs maximaux sont ceux pour lesquels $L = L'$. Si bien que, si n est le degré sur Z d'un sous-corps commutatif maximal de K, on a $[K : Z] = n^2$: le degré d'un corps non commutatif sur son centre est toujours un carré. C'est bien ce qu'on vérifie dans le cas du corps **H** des quaternions où **C** est un sous-corps commutatif maximal $[\mathbf{H} : \mathbf{R}] = 4 = 2^2 = [\mathbf{C} : \mathbf{R}]^2$.

Signalons enfin que R. Brauer a pu munir l'ensemble Br(Z) des classes d'isomorphisme de corps de centre Z et de degré fini sur Z d'une structure de groupe. Ce groupe, appelé le *groupe de Brauer* de Z, reflète un grand nombre de propriétés arithmétiques du corps Z, ainsi que l'ont montré H. Hasse et R. Brauer lui-même.

ROBERT GERGONDEY et *E.U.*

Bibliographie

E. Artin, *Galois Theory*, Notre Dame (Ind.), 1959 / A. Blanchard, *Les Corps non commutatifs*, P.U.F., Paris, 1972 / N. Bourbaki, « Corps commutatifs », in *Éléments de mathématique*, liv. II : *Algèbre*, chap. v, Masson, Paris, 3e éd., 1981 / J. C. Carrera, *Théorie des corps : la règle et le compas*, nouv. éd. Hermann, 1989 / L. Gaal, *Classical Galois Theory*, Chelsea Publ., New York, 4e éd. 1988 / I. Kaplansky, *Fields and Rings*, Univ. of Chicago Press, Chicago, 2e éd. 1972 / C. Mutafian, *Équations algébriques et théorie de Galois*, Vuibert, 1980 / J. Rothman, *Galois Theory*, Springer Verlag, New York, 1990 / P. Samuel, *Théorie algébrique des nombres*, *ibid.*, 2e éd., 1971.

COURBES ALGÉBRIQUES

En fondant la géométrie analytique, Descartes avait substitué au plan de la géométrie d'Euclide l'ensemble $\mathbf{R}^2$ des couples de nombres réels et, de ce fait, à la notion de courbe, celle d'équation. La construction d'un point, puis la détermination d'un lieu géométrique se trouvaient ainsi remplacées par une représentation paramétrique, et une élimination. L'existence de méthodes canoniques d'élimination en théorie des polynômes est sans doute à l'origine de l'intérêt porté aux courbes algébriques, c'est-à-dire, grosso modo, à l'ensemble des points d'un plan où s'annule un polynôme.

Le rôle important de l'homogénéité dans la théorie des polynômes, aperçu au moment où s'élaborait la géométrie projective, a conduit à concevoir les modèles de courbes algébriques comme appartenant au plan projectif, qui a l'avantage d'être compact. D'autre part, si la conception initiale de la géométrie analytique était essentiellement une question de variables réelles, les géomètres algébristes ont été amenés à prendre comme corps de base le corps complexe à cause de la propriété fondamentale suivante : Tout polynôme de degré n à coefficients complexes a exactement n racines complexes, en tenant compte de leur ordre de multiplicité (propriété de clôture algébrique du corps $\mathbf{C}$ des nombres complexes) ; les courbes algébriques ont été les premiers exemples de variétés analytiques complexes.

Cela n'empêche pas, au moins dans le cas des polynômes à coefficients réels, pour aider l'imagination, de s'intéresser à la courbe réelle, lieu des points à coordonnées réelles, dans un plan affine ou même métrique déduit du plan projectif en spécialisant une droite à l'infini et en munissant le repère des propriétés adéquates (orthogonalité, normes) : c'est la raison pour laquelle les mathématiques ont connu toute une abondante « flore » de courbes algébriques remarquables.



1. Courbes irréductibles

Considérons donc un plan projectif complexe, dans lequel les coordonnées homogènes sont x, y, z, et un polynôme (à coefficients réels ou complexes) homogène F de degré n ;

$$F(x, y, z) = 0$$

est l'équation d'une courbe algébrique. Le point A (a, b, c) appartient à la courbe si :

$$F(a, b, c) = 0.$$

Si l'on représente le même point par les coordonnées λa, λb, λc, où $\lambda \neq 0$, on a :

$$F(\lambda a, \lambda b, \lambda c) = \lambda^n F(a, b, c) = 0 ;$$

on voit que ce fait est bien indépendant du choix des coordonnées homogènes et que des polynômes proportionnels définissent la même courbe.

Si $F(x, y, z)$ n'est pas décomposable sur le corps complexe en un produit de facteurs non constants, tout polynôme qui s'annule partout où s'annule F est de la forme :

$$F(x, y, z)G(x, y, z) ;$$

par suite, si on appelle *courbe irréductible* l'ensemble des points où s'annule un polynôme indécomposable, la connaissance de la courbe entraîne celle de son équation (à un facteur constant non nul près).

La décomposition d'un polynôme quelconque en produit de facteurs irréductibles montre alors qu'une courbe algébrique générale doit être considérée comme constituée de *composantes irréductibles*, affectées chacune d'un exposant qui est un entier naturel, sa *multiplicité*. C'est ainsi que la courbe :

$$(x^2 + y^2 - z^2)z^2 = 0$$

est formée de la droite *double* $z = 0$ et de la conique *simple* $x^2 + y^2 - z^2 = 0$.

Bien entendu, la notion de composante irréductible (et la multiplicité correspondante) sont des notions projectives, indépendantes du choix du repère, mais qui sont essentiellement liées au fait que le corps de base est algébriquement clos.

2. Tangentes

Intersection avec une droite

Considérons une droite projective joignant les point $A(x_1, y_1, z_1)$ et $B(x_2, y_2, z_2)$ et son intersection avec la courbe $F(x, y, z) = 0$. On obtient :

$$\begin{aligned} &F(\lambda A + \mu B) \\ &\equiv \lambda^n F(A) + \lambda^{n-1}\mu P_1(A, B) + \lambda^{n-2}\mu^2 P_2(A, B) \\ &\qquad + \dots + \mu^n F(B) = 0. \end{aligned}$$

Si tous les coefficients sont nuls, cette équation est une identité : tout point de la droite appartient à la courbe qui admet la droite comme composante irréductible.

Toute droite qui n'est pas composante de la courbe la coupe en n points (n étant le degré de F) compte tenu de leur ordre de multiplicité, et cet énoncé a exactement la même signification que l'affirmation : une équation algébrique de degré n admet n racines.

Supposons maintenant que A est un point de la courbe $F(A) = 0$, et faisons varier B arbitrairement ; si le polynôme :

$$P_1(A, M) = xf'_{x1} + yf'_{y1} + zf'_{z1}$$

n'est pas nul quel que soit M, toute droite passant par A coupe la courbe en ce point avec la multiplicité 1, à l'exception de la droite :

$$xf'_{x1} + yf'_{y1} + zf'_{z1} = 0,$$

qui coupe la courbe en A avec une multiplicité au moins égale à 2. Le point A est alors appelé un *point simple* de la courbe, et la sécante exceptionnelle est appelée la *tangente* en A (en accord avec les formules différentielles de la géométrie analytique).

Lorsque :

$$P_1(A, M) = \dots = P_{k-1}(A, M) \equiv 0$$

sont nuls quel que soit M, sans qu'il en soit ainsi de P_k, toute droite passant par A coupe la courbe en ce point avec la multiplicité k, à l'exception des droites qui vérifient $P_k(A, M) = 0$, qui coupent la courbe en A avec une multiplicité au moins égale à $k + 1$. Le point A est alors appelé un *point multiple k-uple* de la courbe, et les sécantes exceptionnelles sont appelées les tangentes en A.

Le point A est multiple k-uple de la courbe si toutes les dérivées d'ordre $k - 1$ de F sont nulles en ce point (et pas toutes les dérivées d'ordre k). Bien entendu, un changement de variables projectif sur (λ, μ) ou sur (x, y, z) montre que les résultats précédents sont indépendants des repères.

Équation tangentielle

La question se pose alors de caractériser une courbe algébrique non plus comme l'ensemble de ses points, mais

comme l'ensemble de ses tangentes. L'équation tangentielle est une condition nécessaire et suffisante entre les nombres u, v, w pour que la droite d'équation projective :

$$ux + vy + wz = 0$$

soit tangente à la courbe. L'élimination de x, y, z entre les relations :

$$F(x,y,z) = 0$$

et $$F'_x/u = F'_y/v = F'_z/w$$

conduit à une équation tangentielle algébrique :

$$G(u,v,w) = 0.$$

Lorsque la courbe est irréductible (et n'est pas une droite), l'un des facteurs irréductibles de G représente l'enveloppe proprement dite, c'est-à-dire l'ensemble des tangentes ; les autres facteurs irréductibles sont linéaires : chacun exprime le passage d'une droite par l'un des points singuliers de la courbe.

3. Quelques exemples

La courbe dont l'équation affine est :

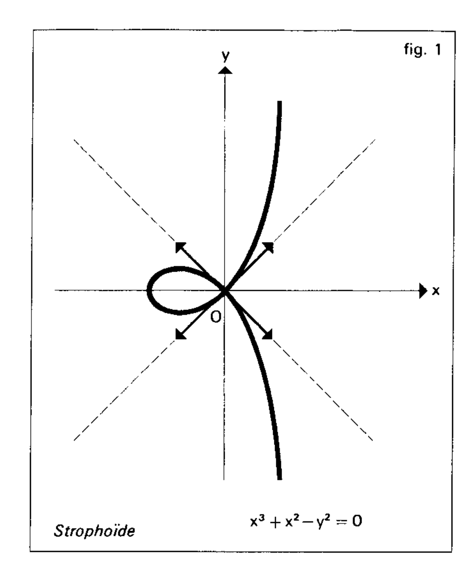
$$x^3 + x^2 - y^2 = 0,$$

c'est-à-dire dont l'équation projective, en coordonnées homogènes, est :

$$x^3 + x^2z - y^2z = 0,$$

est appelée *cubique nodale* (fig. 1).

Elle admet l'origine ($x = 0$, $y = 0$, $z = 1$) comme point double (de multiplicité 2) ; les deux tangentes en ce point sont les droites $y = \pm x$. Le point ($x = 0$, $y = 1$, $z = 0$) est un point simple, pour lequel la tangente $z = 0$ coupe la courbe avec la multiplicité 3 : on l'appelle un *point d'inflexion*. Les cubiques nodales sont toutes projectivement identiques. On rencontre le modèle métrique :



$$y = x\sqrt{(a + x)/(a - x)},$$

qui a quelques propriétés liées à la géométrie du cercle et qu'on appelle la *strophoïde*. On rencontre également le modèle métrique :

$$\begin{cases} x = 3t - t^3 \\ y = 1 - 3t^2 \end{cases}$$

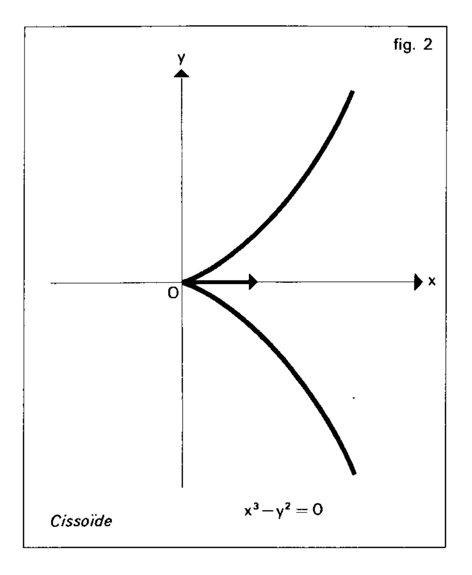
appelé *cubique de Tschirnhausen*, pour lequel la longueur de l'arc est fonction rationnelle des coordonnées.

La courbe dont l'équation affine est :

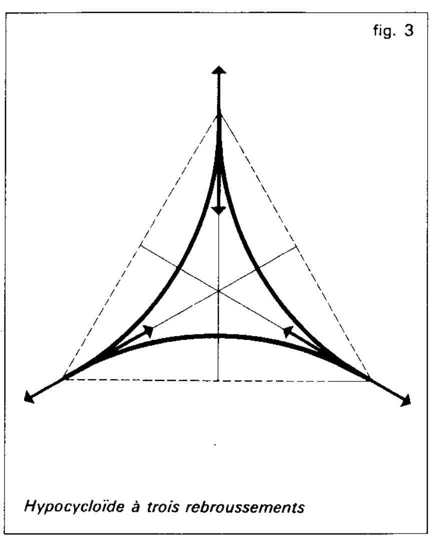
$$x^3 - y^2 = 0$$

est appelée *cubique cuspidale* (fig. 2). Elle admet l'origine ($x = 0$, $y = 0$, $z = 1$) comme point double (de multiplicité 2) ; il y a en ce point une tangente unique $y = 0$. Les cubiques cuspidales sont toutes projectivement identiques. On rencontre le modèle métrique :

$$x^3 + (x - 2R)y^2 = 0$$



Cissoïde



Hypocycloïde à trois rebroussements

lié à la géométrie du cercle, et qu'on appelle la *cissoïde*. La développée de parabole est aussi une cubique nodale.

La courbe dont l'équation projective est :

$$y^2z^2 + z^2x^2 + x^2y^2 - 2xyz(x + y + z) = 0$$

est une *quartique tricuspidale*. Elle admet les trois sommets du repère (0, 0, 1 ; 0, 1, 0 ; 1, 0, 0) comme points doubles, en chacun desquels il y a une tangente unique : respectivement les droites $x = y$, $z = x$, $y = z$, concourantes au point unitaire. Lorsque le repère est choisi suivant un triangle équilatéral dont le point unitaire est le centre, la courbe métrique ainsi définie est connue sous le nom d'*hypocycloïde à trois rebroussements* (fig. 3).

4. Intersection de courbes algébriques

L'étude de l'intersection de deux courbes algébriques F et G de degrés respectifs m et n, qui n'ont aucune composante commune, a été faite par Bezout : il y a un nombre fini de points communs, et chacun est affecté d'un entier naturel, sa multiplicité (d'intersection) ; dénombrés avec cet élément de pondération, il y a mn points communs (sur le corps complexe). Ce théorème, tiré d'une étude attentive du résultant, n'a pas été apprécié à sa juste valeur par les non-spécialistes : faute d'avoir une définition explicite de la multiplicité d'intersection, ils voyaient dans le théorème de Bezout une espèce d'affirmation alchimique. On s'est borné à utiliser le théorème de Bezout dans certains cas simples : lorsque tous les points communs à F et G sont simples sur chacune d'elles, avec des tangentes distinctes, il y a exactement mn points communs. C'est ainsi que la courbe formée de m droites parallèles à Oy et la courbe formée de n droites parallèles à Ox se coupent aux mn sommets d'un quadrillage.

Un autre cas assez simple est celui où un point commun étant multiple d'ordre r pour F et d'ordre s pour G, il n'y a aucune

droite commune aux deux faisceaux de tangentes en ce point : la multiplicité d'intersection est alors rs (et elle est supérieure dans le cas contraire).

La situation n'est devenue claire que lorsque G. Halphen eut montré comment, par une étude locale des courbes F et G en un point commun, on pouvait définir la multiplicité d'intersection.

Considérons par exemple les deux cubiques :

$$F = x^3 - y^2 = 0,$$
$$G = x^3 + y^3 - 2y^2 = 0.$$

Elles ont en commun trois points simples, avec la multiplicité 1, dont les coordonnées satisfont à : $x^3 = 1$, $y = 1$, et le point O qui est pour chacune d'elles un point double ($r = s = 2$) avec une seule tangente Ox. Ce point a la multiplicité d'intersection 6.

On le vérifie en portant $x = t^2$, $y = t^3$, représentation paramétrique de F, dans G :

$$G(t^2, t^3) = t^9 - t^6 = (t^3 - 1)t^6 = 0.$$

5. Étude locale d'un point singulier

Un point d'une courbe algébrique étant pris comme origine des coordonnées dans un modèle affine, l'étude du voisinage de O a été poursuivie par deux méthodes. Celle de Noether consiste à effectuer des transformations birationnelles ayant O pour point d'indétermination : elle relève des techniques de la géométrie algébrique. Celle de Enriques consiste à utiliser les développements de Puiseux : elle relève de l'analyse classique des fonctions d'une variable complexe.

La courbe (irréductible) $F(x, y) = 0$, qui passe en O, définit y comme fonction algébrique de la variable complexe x, multiforme, dont certaines déterminations s'annulent pour $x = 0$.

On appelle *branche algébroïde* de la courbe F un ensemble de ces déterminations qui subit une permutation circulaire lorsque la variable complexe x décrit, dans le plan complexe, un petit cercle autour de l'origine. Si k est le nombre de déterminations constituant une branche, en faisant le changement de variable $x = t^k$, chacune de ces déterminations devient une fonction *entière* de t : $y = f(t)$, et on passe de l'une à la suivante en changeant t en ζt, le coefficient ζ étant une racine primitive k-ième de l'unité.

Les propriétés arithmétiques des exposants qui figurent dans la série entière $f(t)$ ont été interprétées topologiquement (variété analytique complexe) et dans la géométrie infinitésimale de la courbe au voisinage de l'origine.

6. Courbes unicursales

Lorsqu'on a obtenu pour une courbe une représentation paramétrique uniforme, on détient un moyen commode pour l'étude de ses propriétés globales. C'est la raison de l'intérêt porté aux courbes unicursales, c'est-à-dire aux courbes qui, en coordonnées affines, admettent une représentation paramétrique rationnelle :

$$x = P(t)/R(t), \quad y = Q(t)/R(t),$$

où P, Q, R sont des polynômes en t.

Les courbes unicursales sont souvent appelées les courbes rationnelles, car il résulte d'un théorème de Lüroth qu'elles sont les transformées birationnelles des droites projectives.

Les coniques (courbes algébriques irréductibles du second degré) sont rationnelles ; elles admettent en effet la forme réduite projective :

$$xz - y^2 = 0,$$

qui conduit à la représentation :

$$x = t^2, \quad y = t\theta, \quad z = \theta^2,$$

où t, θ sont des paramètres complexes. De ce fait, on peut définir sur une conique le birapport de quatre points, les divisions homographiques et involutives. Ces notions peuvent être étendues à toute courbe rationnelle.

L'équation de la tangente au point courant d'une courbe paramétrique et la théorie des enveloppes montrent qu'il y a identité entre les courbes qui sont rationnelles du point de vue ponctuel et les courbes qui sont rationnelles du point de vue tangentiel.

Les cubiques rationnelles sont les cubiques à point double, dont nous avons donné les deux modèles ; la cubique nodale citée ci-dessus admet pour représentation :

$$x = 4t/(1-t)^2, \quad y = 4t(1+t)/(1-t)^3,$$

et la condition nécessaire et suffisante pour que trois points de la courbe soient alignés est $t_1t_2t_3 = 1$.

La cubique cuspidale citée ci-dessus admet la représentation :

$$x = t^2, \quad y = t^3,$$

et la condition d'alignement de trois points est :

$$1/t_1 + 1/t_2 + 1/t_3 = 0.$$

Les courbes rationnelles sont, parmi les courbes irréductibles de leur degré, celles qui ont les singularités les plus importantes, soit par leur nombre, soit par la complexité de leur structure : si une courbe rationnelle de degré n n'a que des points doubles de même nature que ceux des cubiques, ces points sont au nombre de :

$$(n-1)(n-2)/2.$$

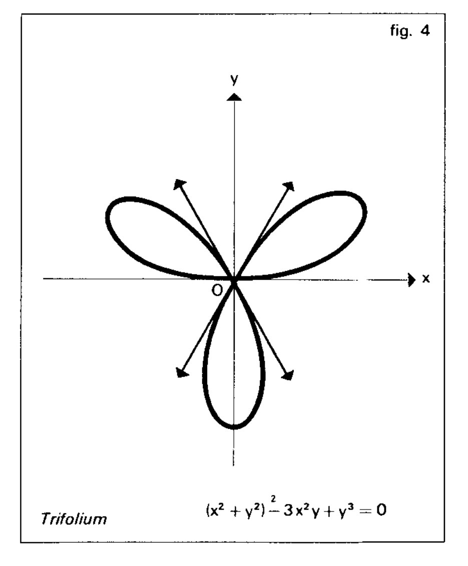
C'est ainsi que la quartique tricuspidale citée plus haut est rationnelle ; elle admet la représentation :

$$x = 1/t^2, \quad y = 1/(t-1)^2.$$

Mais la quartique d'équation (affine) :

$$(x^2 + y^2)^2 - 3x^2y + y^3 = 0,$$

que l'on appelle parfois *trifolium* (fig. 4), admet un seul point singulier, l'origine, qui est un point triple. En coupant cette courbe



Trifolium $(x^2 + y^2)^2 - 3x^2y + y^3 = 0$

par les droites issues de O, on obtient sans difficulté la représentation paramétrique :

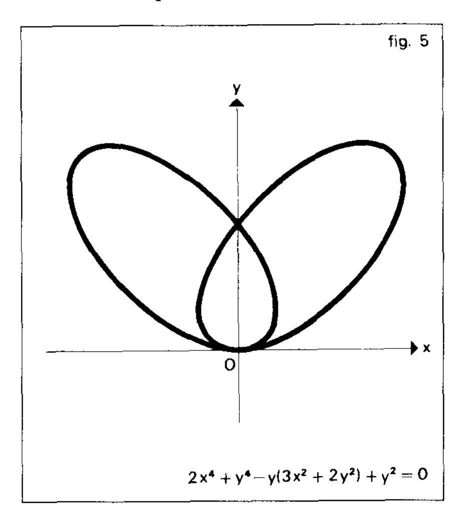
$$x = \frac{3t - t^3}{(1+t^2)^2}, \qquad y = \frac{3t^2 - t^4}{(1+t^2)^2}.$$

L'existence de points doubles plus complexes (points infiniment voisins, contact des branches algébroïdes) permet de donner des exemples d'une nature différente.

La quartique (fig. 5) d'équation (affine) :

$$2x^4 + y^4 - y(3x^2 + 2y^2) + y^2 = 0$$

admet deux points doubles ; le point A



($x = 0$, $y = 1$), qui est un point nodal, et l'origine, qui est un point tacnodal (contact de deux branches algébroïdes). Cette quartique admet la représentation paramétrique :

$$x = t(t^2 - 1)\sqrt{3}/\mathrm{D}, \quad y = t^2/\mathrm{D},$$

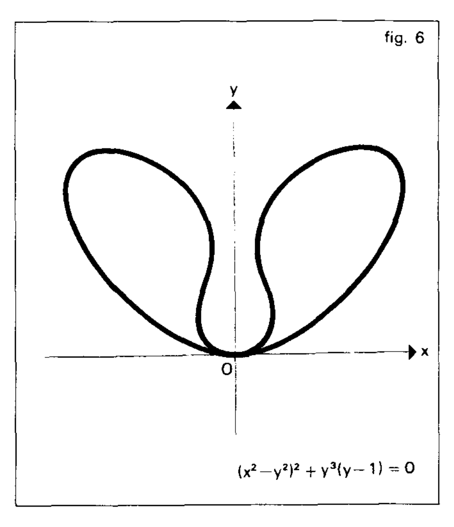
où : $\quad \mathrm{D} = 3t^4 - 8t^2 + 6.$

La quartique (fig. 6) d'équation (affine) :

$$(x^2 - y)^2 + y^3(y - 1) = 0$$

admet un point double unique, à l'origine ; c'est un point oscnodal (osculation de deux branches algébroïdes). Cette singularité suffit à assurer la rationalité, et la quartique proposée admet la représentation paramétrique :

$$x = 4t(t^2 + 3)/\mathrm{D}, \quad y = 16t^2/\mathrm{D},$$
$$\mathrm{D} = (t + 3)^2(t - 1)^2 + 16t^2.$$



7. Courbes elliptiques

Définitions

Nous avons dit que les cubiques sont rationnelles lorsqu'elles ont un point double. Les cubiques sans point singulier sont projectivement réductibles à la forme :

$$y^2 = 4x^3 - g_2x - g_3$$

(dans laquelle l'équation $y = 0$ doit avoir trois racines simples). Cette forme réduite, définie à une homothétie près, dépend du seul paramètre :

$$\mathrm{I} = g_2^3/g_3^2.$$

La définition de la fonction elliptique $\wp(u)$ de Weierstrass met en évidence la représentation paramétrique $x = \wp(u)$, $y = \wp'(u)$; c'est la raison pour laquelle les cubiques sans singularité sont appelées cubiques elliptiques.

L'argument :

$$u = \int \frac{dx}{y} = \int \frac{dx}{\sqrt{4x^3 - g_2x - g_3}}$$

est l'*intégrale abélienne* (de première espèce) attachée à la courbe. Plus généralement, deux fonctions elliptiques de mêmes périodes sont liées par une relation algébrique : elles constituent la représentation paramétrique d'une courbe algébrique dite *courbe elliptique*.

Si ω et ω' sont deux périodes de base d'une fonction elliptique, on appelle parallélogramme de période tout parallélogramme admettant pour sommets les images des nombres complexes :

$$k\omega + k'\omega', (k+1)\omega + k'\omega',$$
$$k\omega + (k'+1)\omega', (k+1)\omega + (k'+1)\omega',$$

où k, k' sont des entiers relatifs. Dans un parallélogramme de période, une fonction elliptique prend le même nombre de fois toute valeur, et la somme des zéros est égale à la somme des pôles. Comme la fonction $\mathfrak{p}(u)$ a un pôle double à l'origine, la fonction :

$$a\mathfrak{p}(u) + b\mathfrak{p}'(u) + c,$$

fournit la condition nécessaire et suffisante d'alignement de trois points d'une cubique elliptique :

$$u_1 + u_2 + u_3 \equiv 0 \quad \text{modulo } \omega, \omega',$$

ω et ω' étant deux périodes de base (cf. FONCTIONS ANALYTIQUES — Fonctions elliptiques et modulaire).

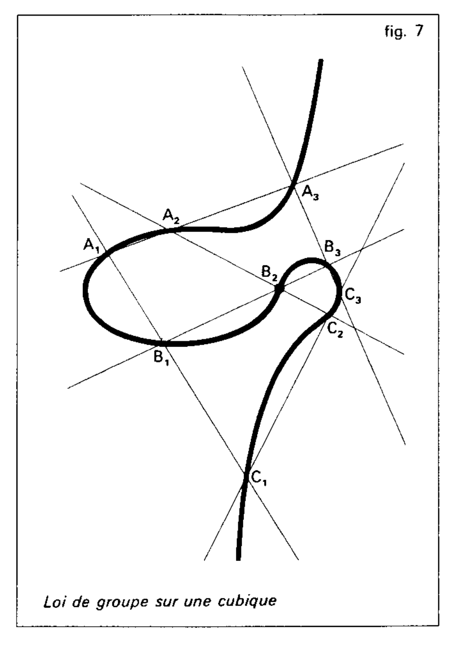
Si, dans l'étude des cubiques nodales, nous faisons le changement de représentation, $u = \log t$ la condition d'alignement devient :

$$u_1 + u_2 + u_3 \equiv 0 \quad \text{modulo } 2i\pi.$$

Pour les cubiques cuspidales, u est le paramètre $1/t$ lui-même. On voit ainsi comment les cubiques rationnelles sont obtenues par dégénérescence des cubiques elliptiques.

Loi de groupe

Le théorème des points alignés (théorème de Lamé) consiste en ceci (fig. 7) : coupons la cubique par trois droites $A_1A_2A_3$, $B_1B_2B_3$, $C_1C_2C_3$, telles que les points $A_1B_1C_1$ et $A_2B_2C_2$, respectivement, soient alignés. Alors $A_3B_3C_3$ sont aussi alignés, car la somme totale des affixes est nulle.



Loi de groupe sur une cubique

Lorsque, dans la condition d'alignement, on fait $u_1 = u_2 = u_3$, on voit que la cubique elliptique a neuf points d'inflexion (la cubique nodale trois et la cubique cuspidale un seul) :

$$u = k\omega/3 + k'\omega'/3$$

(où k, k' sont des entiers relatifs modulo 3). Ces points sont tels que toute droite qui en joint deux en contient un troisième, et cette propriété définit complètement la configuration. Un modèle métrique est obtenu en coupant un triangle équilatéral par la droite de l'infini et le cercle de rayon nul qui lui est concentrique.

Sur la figure 7, qui illustre le théorème des points alignés, nous avons placé B_2 en un point d'inflexion. Cela va nous permettre de montrer la *structure de groupe abélien* de la cubique elliptique, qui a B_2 pour élément neutre : deux points (tels que B_1 et B_3) alignés sur B_2 sont opposés ; quand trois points sont alignés chacun est opposé à la somme des deux autres. Ainsi, la somme $A_1 + A_3$ est C_2 opposé de A_2. Cette opération est évidemment commutative et le théorème de Lamé établit son associativité :

$$(A_2 + A_1) + B_1 = (-A_3) + (-B_3) = C_3,$$
$$A_2 + (A_1 + B_1) = (-C_2) + (-C_1) = C_3.$$

La détermination du point (X, Y) somme des points (x_1, y_1) et (x_2, y_2) se fait rationnellement et cette propriété est en liaison directe avec le théorème d'addition pour la fonction $\wp(u)$:

$$\begin{cases} X = \frac{1}{4}\left(\frac{y_2 - y_1}{x_2 - x_1}\right)^2 - x_1 - x_2 \\ Y = \frac{x_1 y_2 - x_2 y_1}{x_2 - x_1} - X\frac{y_2 - y_1}{x_2 - x_1}. \end{cases}$$

D'un point arbitraire d'une cubique elliptique, on peut lui mener quatre tangentes, dont les contacts ont des arguments qui diffèrent d'une demi-période. D'après un théorème de Salmon, lorsque le point parcourt la cubique, le faisceau de ces quatre tangentes reste projectivement constant.

L'invariant (birapport symétrisé) de ce faisceau s'exprime au moyen de I, ou du quotient ω'/ω des périodes : c'est la signification géométrique de la fonction modulaire :

$$J = \frac{4\,(1 - r + r^2)^3}{27\;r^2(1-r)^2} = \frac{g_2^3}{g_2^3 - 27g_3^2}$$

Les méthodes utilisées pour l'étude des courbes elliptiques ont été généralisées aux courbes qui admettent une intégrale abélienne hyperelliptique ; pour cette raison on les appelle courbes hyperelliptiques. L'extension aux courbes algébriques générales de la méthode paramétrique nécessite l'emploi des fonctions fuchsiennes introduites par Henri Poincaré.

8. Le genre

L'étude locale a permis de définir en chaque point d'une courbe algébrique une ou plusieurs branches algébroïdes : on appelle *place* la donnée d'un point et d'une branche algébroïde issue de ce point. L'ensemble des places d'une courbe est la riemannienne de cette courbe et on appelle *cycle* (parfois aussi *diviseur*) de la courbe une combinaison linéaire formelle à coefficients entiers, positifs, négatifs ou nuls, des points de la riemannienne, un nombre fini seulement de points ayant un coefficient non nul. Les cycles d'une courbe forment un groupe abélien.

On appelle *ordre* d'un cycle la somme de ses coefficients. Un cycle est dit *effectif* (ou positif) si tous ses coefficients sont positifs ou nuls. Un cycle effectif ayant une signification géométrique simple peut, par exemple, être obtenu en envisageant, sur une courbe C irréductible coupée par une courbe algébrique γ, chaque branche algébroïde affectée de la multiplicité de Bezout correspondante.

Plus généralement, étant donné une fraction rationnelle $N(x, y, z)/D(x, y, z)$, où N et D sont deux polynômes homogènes de même degré, dont aucun n'est nul sur toute la courbe C, on peut lui associer le cycle $Z_N - Z_D$, différence des cycles associés au numérateur et au dénominateur : on vérifie en effet que toutes les fractions, formellement différentes, qui ont

la même valeur le long de C, conduisent au même cycle. Les cycles associés aux fractions rationnelles sont d'ordre zéro et forment un sous-groupe abélien.

Deux cycles sont équivalents si leur différence est un cycle associé à une fraction rationnelle. On appelle *série linéaire complète* l'ensemble des cycles effectifs équivalents à un cycle effectif donné ; cet ensemble a la structure d'un espace projectif : si n est l'ordre commun à tous ces cycles et r la dimension de l'espace projectif qu'ils constituent, on désigne la série linéaire par g_n^r. La classe d'équivalence d'un cycle est souvent appelée série linéaire virtuelle.

Deux séries linéaires étant données, prenons un cycle effectif (respectivement Z, Z') dans chacune d'elles : le cycle $Z + Z'$ est effectif et définit une série linéaire complète, somme des deux séries linéaires données. Par exemple, les droites qui coupent une cubique elliptique en un point fixe découpent sur la courbe une g_2^1. Deux g_2^1 étant définies par les droites issues des points A et B de la courbe, la somme est la g_4^3 découpée sur la courbe par les coniques qui passent par A et B ou par tout couple de points équivalents.

Le théorème du reste de Brill-Noether énonce que tous les cycles effectifs (s'il en existe) obtenus par différence des cycles de deux séries linéaires forment une série linéaire. Il a trouvé de nombreuses applications à l'étude des intersections complètes. C'est ainsi, par exemple, que si, parmi les neuf points communs à deux cubiques, il y en a six qui sont situés sur une même conique, les trois autres sont alignés.

Introduit par Riemann, le genre p d'une courbe algébrique irréductible est le nombre des intégrales abéliennes de première espèce, attachées à la courbe, linéairement indépendantes. Les différentielles holomorphes sur la courbe définissent des cycles équivalents qui constituent la classe canonique K : les cycles canoniques effectifs constituent la série canonique qui a l'ordre $2p - 2$ et la dimension $p - 1$. Sur une courbe rationnelle, K a l'ordre -2. Sur une courbe elliptique, K a l'ordre zéro ; seul le cycle nul appartient à la série canonique qui a la dimension zéro. Cela tient à ce que, selon un théorème de Liouville, les seules fonctions elliptiques holomorphes sont les constantes.

Soit G un cycle d'une série linéaire g_n^r : si la classe $K - G$ contient des cycles effectifs, on dit que la série est spéciale, et le nombre de cycles effectifs de $K - G$ linéairement indépendants est appelé son indice de spécialité i. Lorsqu'il n'y en a pas (ce qui est, par exemple, le cas pour $n > 2p - 2$), la série g_n^r est dite régulière.

La signification géométrique du genre est alors donnée par les deux théorèmes suivants : Si une série linéaire complète g_n^r est régulière, on a $r = n - p$ (Riemann). Plus généralement, si une série linéaire complète g_n^r a l'indice de spécialité i, on a $r = n - p + i$ (Riemann-Roch).

L'ensemble des séries linéaires complètes g_p^i d'ordre p peut, par application du théorème du reste, être muni d'une structure de variété abélienne de dimension p : c'est la jacobienne de la courbe.

La classe canonique et, par conséquent, le genre p ont été introduits d'une façon purement algébrique par Enriques, au moyen d'une construction tirée du jacobien. Liée à la théorie des enveloppes, cette construction élégante perd malheureusement sa valeur en géométrie algébrique abstraite sur les corps de caractéristique non nulle.

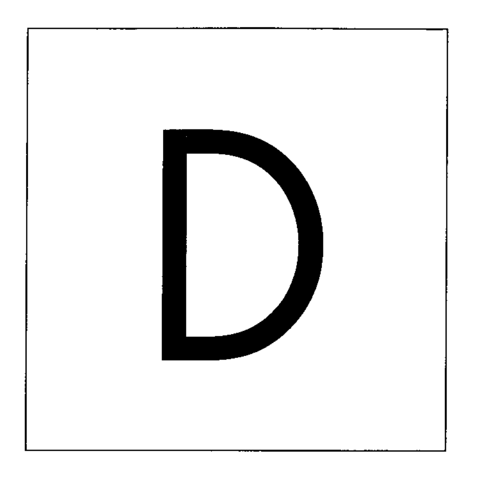
Une autre définition du genre a été présentée par Weierstrass, qui s'intéresse aux cycles associés aux pôles d'une fraction

rationnelle définie sur une courbe irréductible. Une place P munie de la multiplicité n peut être l'unique pôle d'une fraction rationnelle, à condition que n ne fasse pas partie d'un certain ensemble de p entiers, les « lacunes », inférieures à $2p - 1$. Si P pris sur la riemannienne est pris hors d'un certain ensemble fini, les lacunes sont $1, 2, ..., p$. En géométrie algébrique abstraite, sur un corps de caractéristique non nulle, la seconde partie du théorème de Weierstrass n'est pas vraie : il peut arriver que pour toute place P les lacunes diffèrent de $1, 2, ..., p$.

LUC GAUTHIER

Bibliographie

A. CHENCINER, *Courbes algébriques planes*, Publications mathématiques de l'université Paris-VII, Paris, 1978 / H. CLEMENS, *A Scrapbook of Complex Curve Theory*, Plenum Press, New York-Londres, 1980 / J. FRESNEL, *Géométrie algébrique*, U.F.R. de mathématiques et d'informatique de l'université de Bordeaux, Talence, 1989 / FULTON, *Algebraic Curves*, Addison-Wesley, New York, 1989 / P. A. GRIFFITHS, *Introduction to Algebric Curves*, A.M.S., Providence (R.I.), 1989 / A. GROTHENDIECK, *Fondements de la géométrie algébrique*, 2 vol. Secrétariat mathématique, 1985 / J. PICHON, *Les Courbes dans le plan et dans l'espace*, Ellipses, 1987 ; *Équations $f(n) = O$, Courbes $y = f(x)$*, Marketing, 1988 / J. VELU, *Courbes elliptiques*, Secrétariat mathématique, 1978.



DÉRIVÉES PARTIELLES
ÉQUATIONS AUX

Les équations aux dérivées partielles sont sans doute le domaine des mathématiques où le lien avec la physique est le plus étroit. Il ne s'agit pas seulement du fait que les recherches les plus actives, et en général les plus importantes, ont été motivées par des questions de physique. Il s'agit aussi, et surtout, du fait que les idées apportées par la physique, et notamment la mécanique, transposées ensuite dans un cadre plus général, ont fourni les outils les plus puissants de leur étude. On en trouvera plusieurs exemples dans les articles qui suivent.

Cette étude a eu, à son tour, une influence fondamentale sur le développement général de l'analyse mathématique au XVIIIe siècle. Le cas le plus célèbre est la question de savoir si toute fonction est développable en série trigonométrique, posée d'abord par les travaux de d'Alembert et Daniel Bernoulli sur l'équation des cordes vibrantes, puis par ceux de Fourier

sur l'équation de la chaleur (cf. SÉRIES TRIGONOMÉTRIQUES). On peut citer aussi la théorie du potentiel (cf. POTENTIEL ET FONCTIONS HARMONIQUES), née de l'étude de la gravitation et du potentiel électrique, et devenue un domaine à part entière de l'analyse mathématique. Les méthodes par dualité, que George Green appliquait déjà aux équations aux dérivées partielles dans la décennie des années 1830, ont joué un rôle essentiel ; elles ont été reprises par Henri Poincaré, puis par Jacques Hadamard. Ce point de vue des solutions faibles des équations aux dérivées partielles à été systématiquement utilisé par l'école russe (Izraïl Gelfand, Sergueï Sobolev) et a amené Laurent Schwartz à l'élaboration de la théorie des distributions (cf. théorie des DISTRIBUTIONS) qui constitue de nos jours le cadre naturel de la théorie des équations aux dérivées partielles linéaires.

La théorie des distributions s'impose aussi pour les problèmes non linéaires, car les données non linéaires génèrent, même à partir de données régulières, des solutions *singulières* (interprétables dans le langage des distributions). Ces idées apparaissent déjà chez Riemann (1860).

L'interaction entre le développement de la physique et de l'analyse fonctionnelle s'est effectuée par l'intermédiaire des équations aux dérivées partielles : l'espace de Hilbert L^2 est l'espace naturel des solutions de l'équation de Schrödinger de la mécanique quantique ; de même, l'espace de Sobolev H^1 (cf. chap. 2 *Le type elliptique*, dans la partie A ci-dessous – Sources et applications, et 6 *Opérations sur les représentations et les approximations*, in représentation et approximation des FONCTIONS) est l'espace naturel des solutions des problèmes de mécanique des milieux continus décrits par des équations elliptiques.

La transformation de Fourier est particulièrement bien adaptée à l'étude des équations à coefficients constants et elle a permis de disposer, dès le début du siècle, d'exemples explicites qui ont servi de modèles. L'étude fine de cette transformation apparaît lorsque l'on veut établir des relations entre la propagation ondulatoire et la propagation corpusculaire (par l'intermédiaire des équations des ondes ou de Schrödinger). Une difficulté fondamentale est la suivante : il n'existe pas de fonction non nulle à support compact dont la transformée de Fourier soit aussi à support compact. Cette observation simple est une version mathématique naïve de l'inégalité de Heisenberg. C'est pour essayer de contourner cette difficulté que Hormander et ses collaborateurs ont introduit les concepts de microlocalisation.

CLAUDE BARDOS et MARTIN ZERNER



A. Sources et applications

On se propose de décrire très sommairement quelques types classiques d'équations aux dérivées partielles issues principalement de la physique et de préciser leurs interventions dans des domaines variés des mathématiques.

Alors que les solutions des équations différentielles ordinaires dépendent d'une ou de plusieurs constantes arbitraires, celles des équations et systèmes d'équations aux dérivées partielles dépendent de fonctions arbitraires ; il y a donc des

familles beaucoup plus riches de solutions. Ce fait se voit sur l'exemple particulièrement simple d'une équation linéaire du premier ordre :

$$\sum_{j=1}^{n} a_j \frac{\partial u}{\partial x_j} + cu = f, \quad j = 1, ..., n;$$

on lui associe le *système différentiel des caractéristiques* :

$$x_j'(s) = a_j(x(s)), \quad j = 1, ..., n,$$

dont les trajectoires sont les *courbes caractéristiques* de l'équation. L'équation aux dérivées partielles équivaut alors à une équation différentielle ordinaire sur chaque courbe caractéristique. Posant :

$$w(s) = u(x(s)),$$

cette équation différentielle s'écrit :

$$w'(s) + c(x(s))w(s) = f(x(s));$$

il faut la compléter par une donnée initiale sur chaque caractéristique, ce qui introduit une fonction arbitraire. On remarquera sur cet exemple qu'une solution d'une équation sans second membre ($f = 0$) ne peut s'annuler en un point sans s'annuler sur toute la courbe caractéristique qui passe par ce point.

La façon la plus courante de déterminer une solution, en particulier dans les problèmes d'origine physique, est de fixer les valeurs de la fonction et d'une ou plusieurs de ses dérivées sur des hypersurfaces. On dit que le problème est *bien posé* lorsque cela détermine une solution et une seule. Pour traduire une situation physique, un problème doit non seulement être bien posé au sens précédent, mais posséder en plus une propriété de stabilité : la solution doit dépendre continûment des données (en un sens à préciser dans chaque problème particulier). Cette condition est automatiquement vérifiée dans les problèmes linéaires (c'est une conséquence du théorème du graphe fermé ; cf. espaces vectoriels TOPOLOGIQUES).

Il est remarquable, ce qui sera évident ci-dessous, que les premiers travaux systématiques ont porté sur des équations du second ordre, qui se sont présentées en mécanique, puis dans la théorie de la chaleur. L'étude des équations du premier ordre, la plus simple du point de vue mathématique, n'est venue que plus tard.

1. L'équation des ondes et le type hyperbolique

L'équation des ondes (équation de d'Alembert) :

$$(1) \quad \frac{\partial^2 u}{\partial t^2} - c^2\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}\right) = 0$$

régit le comportement de la densité dans une onde sonore, c'est-à-dire une perturbation de faible amplitude d'un gaz non visqueux au repos. Dans une série de phénomènes physiques représentés par des grandeurs vectorielles, chaque composante des vecteurs concernés obéit à cette même équation : ondes transversale et longitudinale dans un solide élastique, ondes électromagnétiques, etc. Il faut y ajouter les phénomènes analogues dépendant seulement d'une ou deux variables d'espace ; parmi eux, les vibrations transversales d'un fil élastique donnent lieu au cas particulier de l'équation des cordes vibrantes :

$$(2) \quad \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0,$$

la plus ancienne à avoir été explicitement étudiée (dans la décennie de 1740 par d'Alembert, Daniel Bernoulli et Euler).

La possibilité d'écrire toutes les solutions de l'équation des cordes vibrantes sous la forme :

$$u(t,x) = f(x - ct) + g(x + ct)$$

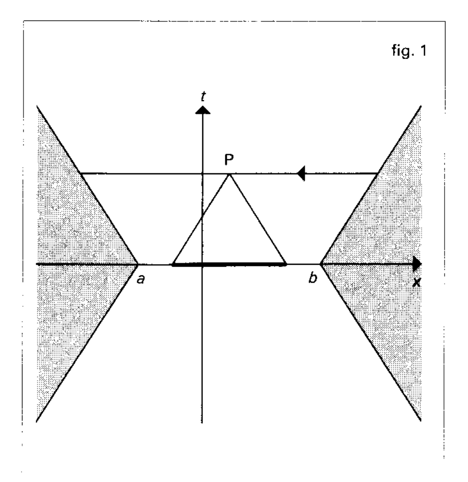
permet d'en voir facilement certaines propriétés :

a) *Le problème de Cauchy* est bien posé, tant dans le futur ($t > t_0$) que dans le passé ($t < t_0$). Ce problème s'énonce ici : « Trouver u vérifiant l'équation (2) et de plus les conditions :

$$u(t_0, x) = u_0(x) \quad \text{et} \quad \frac{\partial u}{\partial t}(t_0, x) = u_1(x),$$

où u_0 et u_1 sont des fonctions données. »

b) *Les solutions se propagent à la vitesse c.* Cette affirmation délibérément vague peut se préciser de plusieurs façons. Par exemple en revenant au problème de Cauchy, avec t_0 nul pour simplifier, si u_0 et u_1 s'annulent en dehors d'un intervalle $[a, b]$, $u(t, x)$ s'annule lorsque x est en dehors de l'intervalle $[a - c|t|, b + c|t|]$ (fig. 1). Sur cette figure, si les données de Cauchy $u(0, x)$ et $\partial u/\partial t\,(0, x)$ s'annulent en dehors de $[a, b]$, la solution u s'annule sur toute la région ombrée.



Propagation des solutions et domaine de dépendance pour l'équation des cordes vibrantes : les pentes des droites obliques sont ± 1/c.

Cette propriété peut s'exprimer de façon équivalente en termes de *domaines de dépendance* : $u(t, x)$ ne dépend que des restrictions de u_0 et u_1 à l'intervalle $[x - c|t|, x + c|t|]$ (fig. 1). Le domaine de dépendance du point P est le segment renforcé sur l'axe des x ; on peut calculer $u(\mathrm{P})$ à l'aide des données de Cauchy sur ce segment.

c) *Les singularités de la solution se propagent, elles aussi, à la vitesse c.* Si par exemple u_1 présente une discontinuité au point x_0, on retrouvera des discontinuités des dérivées premières de u aux points $(t, x - ct)$ et $(t, x + ct)$.

La possibilité d'expliciter toutes les solutions sous une forme simple est tout à fait spéciale à l'équation des cordes vibrantes. Mais les propriétés que nous en avons tirées se démontrent par d'autres méthodes pour toute une classe d'équations et de systèmes qu'on appelle hyperboliques. C'est dans cette classe qu'on trouve les équations de base auxquelles obéissent les phénomènes physiques réversibles, à commencer par l'équation des ondes. La formulation du problème de Cauchy n'a pas besoin d'être changée, du moins tant qu'on se limite à une équation du second ordre. La géométrie de la propagation devient plus compliquée et, pour la décrire dans le cas général, même linéaire, il faut recourir aux courbes bicaractéristiques. Nous allons

décrire ici deux résultats simples sur l'équation :

$$(3) \quad \frac{\partial^2 u}{\partial t^2} - \operatorname{div}_x (c^2(x) \operatorname{grad}_x u) + b \frac{\partial u}{\partial t} = 0,$$

équation qui décrit la propagation d'une onde amortie avec une vitesse dépendant du point où on se trouve (comme le son dans une atmosphère inhomogène). Nous poserons :

$$\sigma(t, x, \tau, \xi) = \tau^2 - c^2(x) \| \xi \|^2$$

(c'est ce qu'on appelle le *symbole principal*).

Soit u une solution de (3) qui s'annule d'un côté d'une hypersurface Σ d'équation : $S(t, x) = 0$. On peut démontrer que si u ne s'annule pas sur tout un voisinage de Σ, S vérifie l'équation aux dérivées partielles non linéaire du premier ordre :

$$(4) \qquad \sigma\left(t, x, \frac{\partial S}{\partial t}, \operatorname{grad}_x S\right) = 0 ;$$

on dit que c'est une *hypersurface caractéristique*.

Supposons de plus que u soit deux fois continûment dérivable en dehors de Σ mais discontinue sur Σ. Nous noterons $[u]$ sa discontinuité ; $[u]$ est donc définie sur Σ comme limite de $u(x)$ quand x tend vers un point de Σ en restant du côté où u ne s'annule pas identiquement. Si on cherche alors à écrire l'équation (3) au sens des distributions, on voit apparaître une double couche portée par Σ de densité $[u]$ que multiplie le premier membre de (4), ce qui donne la démonstration dans ce cas particulier. Il apparaît également une simple couche dont la densité est à un coefficient non nul près :

$$\frac{\partial S}{\partial t} \frac{\partial [u]}{\partial t} - c^2(x) \operatorname{grad}_x S \, . \operatorname{grad}_x [u] + h [u],$$

où h est une fonction qui se calcule à partir des coefficients de (3) et de S. On voit donc que la discontinuité $[u]$ vérifie une équation linéaire du premier ordre dont le système caractéristique est :

$$(5) \begin{cases} T'(s) = \dfrac{\partial S}{\partial t} = \dfrac{1}{2} \dfrac{\partial \sigma}{\partial \tau}\left(t, x, \dfrac{\partial S}{\partial t}, \operatorname{grad}_x S\right) \\ X'(s) = -c^2(x) \operatorname{grad}_x S \\ \qquad = \dfrac{1}{2} \operatorname{grad}_\xi \sigma\left(t, x \dfrac{S}{t}, \operatorname{grad}_x S\right) ; \end{cases}$$

les trajectoires de ce système qui ont un point sur Σ sont tout entières sur cette hypersurface du fait qu'elle est caractéristique, ce sont des *courbes bicaractéristiques* de l'équation (3).

Le point important est que u vérifie une équation différentielle le long de chacune de ces bicaractéristiques, d'où il résulte que sa valeur en un point la détermine sur toute la bicaractéristique issue de ce point. C'est en ce sens que la discontinuité se propage le long des bicaractéristiques. De même, si à l'instant $t = 0$ la discontinuité présente un pic au voisinage de x_0, ce pic se retrouvera en chaque point de la bicaractéristique issue de $(0, x_0)$. On notera la nature cinématique des bicaractéristiques : elles représentent des points qui se déplacent à la vitesse c perpendiculairement aux surfaces $S = C^{te}$; on retrouve ainsi le comportement des rayons lumineux.

Pour achever de se ramener à la définition générale des bicaractéristiques, on vérifiera que, si on pose :

$$\tau(s) = \frac{\partial S}{\partial t}(T(s), X(s)),$$

$$\xi(s) = \operatorname{grad}_x S(T(s), X(s)),$$

on a sur les bicaractéristiques :

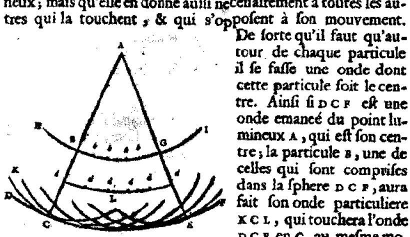
$$\tau'(s) = -\frac{1}{2} \frac{\partial \sigma}{\partial t} = 0, \quad \xi'(s) = -\frac{1}{2} \operatorname{grad}_x \sigma.$$

Il est remarquable que les conclusions de l'étude que nous venons de présenter aient été dégagées par Huygens dans son *Traité de la lumière* (écrit en 1678 et publié en 1690) sur la base de considérations physiques et géométriques avant que qui que ce soit n'ait écrit une équation aux dérivées partielles (cf. fig. 2).

DE LA LUMIERE. CHAP. I. 17

du corps lumineux une infinité d'ondes, quoique issuës de points differens de ce corps, s'unissent en sorte que sensiblement elles ne composent qu'une onde seule, qui par consequent doit avoir assez de force pour se faire sentir. Ainsi ce nombre infini d'ondes qui naissent en mesme instant de tous les points d'une étoile fixe, grande peut estre comme le Soleil, ne font sensiblement qu'une seule onde, laquelle peut bien avoir assez de force pour faire impression sur nos yeux. Outre que de chaque point lumineux il peut venir plusieurs milliers d'ondes dans le moindre temps imaginable, par la frequente percussion des corpuscules, qui frappent l'Ether en ces points; ce qui contribuë encore à rendre leur action plus sensible.

Il y a encore à considerer dans l'émanation de ces ondes, que chaque particule de la matiere, dans laquelle une onde s'etend, ne doit pas communiquer son mouvement seulement à la particule prochaine, qui est dans la ligne droite tireé du point lumineux; mais qu'elle en donne aussi necessairement à toutes les autres qui la touchent, & qui s'opposent à son mouvement. De sorte qu'il faut qu'autour de chaque particule il se fasse une onde dont cette particule soit le centre. Ainsi si DCF est une onde emaneé du point lumineux A, qui est son centre; la particule B, une de celles qui sont comprises dans la sphere DCF, aura fait son onde particuliere KCL, qui touchera l'onde DCF en C, au mesme moment que l'onde principale, emanée du point A, est parve-



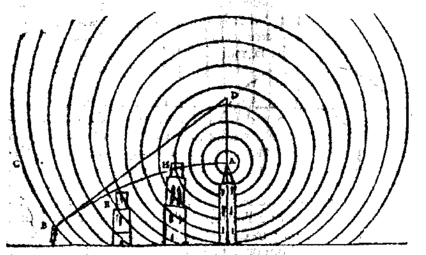
C nuë

Une série de problèmes importants dans les applications concernent la diffusion (« scattering ») : c'est l'étude du rapport entre les comportements asymptotiques des solutions pour t tendant vers moins et plus l'infini.

On appelle systèmes d'évolution du premier ordre symétriques, ceux qui ont la

44 TRAITE

vent s'étendre plus amplement vers en haut, & moins vers en bas, mais vers les autres endroits plus ou moins selon qu'ils



approchent de ces deux extremes. Ce qui estant, il s'ensuit necessairement que toute ligne, qui coupe une de ces ondes à angles droits, passe au dessus du point A, si ce n'est la seule qui est perpendiculaire à l'horizon.

Soit BC l'onde qui porte la lumiere au spectateur qui est en B, & que BD soit la droite qui coupe cette onde perpendiculairement. Or parce que le rayon ou la ligne droite, par laquelle nous jugeons l'endroit où l'objet nous paroit, n'est autre chose que la perpendiculaire à l'onde qui arrive à nostre oeil, comme l'on peut entendre par ce qui à esté dit cy dessus, il est manifeste que le point A s'appercevra comme estant dans la droite BD, & ainsi plus haut qu'il n'est en effet.

De mesme si la Terre est AB, & l'extremité de l'Atmosphere

CD;

Une méthode importante est l'approximation des hautes fréquences qui repose sur la construction de solutions approchées de la forme $\exp(i\nu\varphi)\,v$, où v possède un développement limité en ν au voisinage de l'infini. On trouve que φ vérifie encore l'équation des caractéristiques (4) et que chaque terme du développement limité de v vérifie une équation linéaire du premier ordre dont les courbes caractéristiques sont des bicaractéristiques de l'équation étudiée. On arrive à donner une base mathématique rigoureuse aux calculs qui se faisaient en optique, dans l'étude des radars, etc.

forme :

$$\frac{\partial u}{\partial t} = \sum_{j=1}^{n} \mathrm{A}_j \frac{\partial u}{\partial x_j} + \mathrm{B}u,$$

où u est la fonction inconnue (vectorielle) et les A_j des matrices autoadjointes. Ce sont des systèmes hyperboliques et, en fait, la plupart des systèmes les plus importants peuvent se mettre sous cette forme.

L'équation de Dirac, qui décrit l'évolution d'une particule relativiste de spin ½, est un tel système, où la fonction d'onde de la particule a quatre composantes. L'équation des ondes peut elle aussi se mettre sous

la forme d'un système symétrique. Pour cela, on prend pour fonctions inconnues :

$$u_0 = \frac{\partial u}{\partial t},$$

$$u_j = \frac{\partial u}{\partial x_j}; \quad j = 1, ..., n.$$

L'équation des ondes devient alors :

$$\frac{\partial u_0}{\partial t} = \sum_{j=1}^{n} \frac{\partial u_j}{\partial x_j},$$

et on complète le système en écrivant les équations qui expriment que les dérivations par rapport aux x_j et à t commutent :

$$\frac{\partial u_j}{\partial t} = \frac{\partial u_0}{\partial x_j}, \quad j = 1, ..., n.$$

Le sytème obtenu est bien symétrique.

D'autres équations que les hyperboliques ont des solutions qui se propagent, on en verra des exemples – tous non linéaires – dans la partie **C** ci-après – Équations non linéaires. Mais, dans ces autres équations, la vitesse de propagation dépend toujours de la solution considérée.

2. Le type elliptique

L'équation de Laplace, ou de Poisson

Si dans l'équation des ondes on s'intéresse à des solutions stationnaires (c'est-à-dire indépendantes du temps), on tombe sur l'équation de Poisson :

$$(6) \qquad \Delta u = \sum_{j=1}^{n} \frac{\partial^2 u}{\partial x_j^2} = f,$$

plus connue sous le nom d'équation de Laplace lorsque le second membre est nul, et prototype des équations elliptiques. De même, si on s'intéresse aux solutions qui ne dépendent du temps que par un facteur $e^{i\omega t}$ ou $\cos \omega(t - t_0)$, on voit qu'elles vérifient l'équation de Helmholtz :

$$\Delta u + k^2 u = 0.$$

Cette équation a des propriétés tout à fait analogues à celle de Laplace.

On retrouve l'équation de Laplace (à deux variables indépendantes) comme conséquence des conditions de Cauchy-Riemann. Elle est donc vérifiée par la partie réelle et la partie imaginaire pure de toute fonction analytique d'une variable complexe. Ce fait a été la source de certains problèmes de la théorie des équations aux dérivées partielles (quelles sont les propriétés des fonctions analytiques qui peuvent être généralisées ici ?). Il a aussi été la source de certaines applications dont la plus célèbre est la méthode de Joukovski (à une certaine approximation, le calcul d'un écoulement incompressible autour d'une aile se ramène à un problème de représentation conforme).

Les problèmes bien posés pour l'équation de Laplace concernent en général les solutions sur un ouvert borné Ω de $\mathbf{R}^n$ dont nous noterons Γ la frontière. Les deux plus usuels sont :

– Le *problème de Dirichlet* : « Trouver u vérifiant (6) sur Ω et dont la restriction à Γ est donnée. »

– Le *problème de Neumann* : « Trouver u vérifiant (6) sur Ω et dont la dérivée normale sur Γ est donnée. »

À vrai dire, ce dernier n'est pas tout à fait bien posé. D'abord il n'y a pas unicité puisqu'en ajoutant à une solution une constante, on retrouve une autre solution. D'autre part, pour qu'il y ait existence, les données doivent vérifier une condition que nous allons trouver en utilisant l'outil

fondamental pour ce genre de questions, la *formule de Green* :

$$\int_\Omega v\,\Delta u\,dx = -\int_\Omega \operatorname{grad} v \,.\, \operatorname{grad} u\,dx + \int_\Gamma v\,\partial_n u\,d\sigma,$$

où ∂_n désigne la dérivée normale sortante et $d\sigma$ la mesure superficielle ; nous notons g la dérivée normale sortante (donnée) de u et nous appliquons la formule de Green avec $v = 1$, ce qui nous donne compte tenu de (6) :

$$\int_\Omega f\,dx = \int_\Gamma g\,d\sigma. \tag{7}$$

Cette condition a une interprétation très claire dans tous les cas où $-\operatorname{grad} u$ est le flux d'une grandeur (par exemple si $-u$ est le potentiel des vitesses d'un liquide animé d'un mouvement irrotationnel ; ou si on est dans le cas stationnaire de l'équation de la chaleur, u est alors la densité d'énergie interne et $-\operatorname{grad} u$, moyennant un choix d'unités, son flux). Dans cette situation, le premier membre de (7) est la quantité de la grandeur en question créée à l'intérieur de Ω et le second membre la quantité qui sort à travers la frontière.

On rencontre très souvent des problèmes mêlés, c'est-à-dire où on donne u sur une partie de la frontière et sa dérivée normale sur le reste. Ces problèmes sont bien posés. Une série de problèmes généralisent celui de Neumann (condition de Newton, dérivées obliques...).

Les solutions des équations de Poisson et de Laplace et des équations analogues possèdent de multiples propriétés : elles sont analytiques, ne peuvent pas avoir de maximum ni de minimum à l'intérieur d'un domaine où l'équation est vérifiée. On appelle *fonctions harmoniques* les fonctions qui vérifient l'équation de Laplace (cf. POTENTIEL ET FONCTIONS HARMONIQUES).

Applications à la topologie

Considérons une variété compacte V et munissons-la d'une métrique riemannienne, comme il est possible de le faire. On sait que cette métrique riemannienne induit pour tout k un produit scalaire sur l'espace vectoriel $\wedge_k$ des champs de formes extérieures de degré k sur V. La différentiation extérieure d qui opère de $\wedge_k$ dans $\wedge_{k+1}$ possède donc un adjoint d^*. Posons :

$$\Delta = dd^* + d^*d.$$

Dans le cas des fonctions sur un espace euclidien, on retrouve bien le laplacien usuel ; c'est essentiellement la formule :

$$\Delta u = \operatorname{div} \operatorname{grad} u.$$

Notons en général Δ_k la restriction de Δ à $\wedge_k$. On a alors le *théorème de Hodge-de Rham* : la dimension du noyau de Δ_k est le k-ième nombre de Betti de V.

Ce théorème a reçu une série de généralisations dont le célèbre théorème de l'indice d'Atiyah-Singer et, tout récemment, le résultat d'Alain Connes classifiant les feuilletages en attachant un indice aux opérateurs elliptiques sur chaque feuillet.

Principe des travaux virtuels et formulations variationnelles

Les équations vérifiées par le déplacement d'un solide élastique en équilibre forment un système elliptique. Nous allons en donner une formulation fondée sur le principe des travaux virtuels.

Soit Ω le volume occupé par le solide au repos, Γ sa frontière. Supposons que sur une partie Γ_0 de Γ le solide soit fixé et que sur le reste Γ_1 on lui applique une force de densité superficielle g. Un déplacement

admissible est un champ de vecteurs qui s'annule sur Γ_0. Soit $u(x)$ le déplacement du point du solide qui se trouve en x au repos. On sait que chaque composante de la densité des forces élastiques s'écrit :

$$\sum_{k=1}^{3} \frac{\partial \sigma_{j,k}}{\partial x_k}$$

σ est le tenseur des contraintes, qui est symétrique. Leur travail dans un déplacement virtuel admissible v est :

$$(8) \quad W_i = \int_\Omega \sum_{j,k} v_j \frac{\partial \sigma_{j,k}}{\partial x_k} dx - \int_\Gamma \sum_{j,k} v_j \sigma_{j,k} n_k \, dS,$$

où l'indice i évoque le mot « interne » ; les n_k sont les composantes de la normale sortante et dS la mesure superficielle sur Γ. Soit directement par des considérations physiques qu'il serait trop long de décrire ici, soit en appliquant à (8) une intégration par parties et en tenant compte de la symétrie de σ, on arrive à la nouvelle formule :

$$(9) \quad W_i = \frac{1}{2} \int_\Omega \sum_{j,k} \left(\frac{\partial v_j}{\partial x_k} + \frac{\partial v_k}{\partial x_j} \right) \sigma_{j,k} \, dx.$$

La loi de comportement donne l'expression de σ qui dépend des dérivées de u et (si le solide étudié est inhomogène) de x. Usuellement, si la déformation est assez petite pour que le solide reste élastique, il suffit de supposer la dépendance en u linéaire. Mais il est important de comprendre que nous n'avons pas besoin de cette simplification ici ; nous ne la faisons donc pas.

Avant d'écrire que le travail virtuel des forces élastiques est égal à celui de la force appliquée, il reste à préciser ce qu'est un déplacement admissible. Il faut qu'il satisfasse aux liaisons imposées au système, c'est-à-dire ici qu'il s'annule sur Γ_0. Il faut de plus qu'il vérifie une condition de régularité qui assure au minimum l'existence de l'intégrale qui figure dans la formule (9), nous y reviendrons. On aboutit à la *formulation variationnelle* du problème de la recherche du déplacement à l'équilibre : Trouver un déplacement admissible u qui vérifie pour tout déplacement admissible v :

$$(10) \quad \frac{1}{2} \int_\Omega \sum_{j,k} \left(\frac{\partial v_j}{\partial x_k} + \frac{\partial v_k}{\partial x_j} \right) \sigma_{j,k}(x, \operatorname{grad} u) \, dx = \int_{\Gamma_1} v \, . \, g \, dS.$$

En généralisant un peu, on arrive à la formulation suivante, où V est un espace de fonctions sur Ω (qui peuvent être à valeurs vectorielles) et les F_i des fonctions connues (il est commode et peu restrictif de supposer qu'elles vérifient $F_i(x, 0, 0) = 0$) : « Trouver u appartenant à V telle que pour toute fonction v appartenant à V :

$$(11) \quad \int_\Omega \left[\sum_{i=1}^{n} \frac{\partial v}{\partial x_i} F_i(x, u, \operatorname{grad} u) + v \, . \, F_0(x, u, \operatorname{grad} u) \right] dx = \int_{\Gamma_1} v \, . \, g \, dS.$$

Des intégrations par parties montrent que u vérifie l'équation aux dérivées partielles du second ordre :

$$(12) \quad \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(x, u(x), \operatorname{grad} u(x)) = F_0(x, u, \operatorname{grad} u)$$

et des conditions aux limites qui dépendent du choix de l'espace V.

Le plus grand avantage des formulations variationnelles est justement de contenir à la fois l'équation aux dérivées partielles et les conditions aux limites.

Voyons cela de plus près pour les problèmes mêlés. Sous forme classique, nous cherchons u deux fois différentiable qui vérifie (6) sur Ω, $u = g_0$ sur Γ_0 et $\partial_n u = g_1$ sur Γ_1, où g_0 et g_1 sont des fonctions données. C'est la formule de Green qui va nous permettre de passer à la forme variationnelle. Elle assure que pour toute fonction v suffisamment régulière et nulle sur Γ_0 :

$$(13) \quad -\int_\Omega (\text{grad}\, v \,.\, \text{grad}\, u + vf)\, dx = \int_{\Gamma_1} v g_1 \, dS.$$

Il reste à définir l'espace V. Débarrassons-nous d'abord d'un détail : une fonction appartenant à V devra s'annuler sur Γ_0 alors que la solution u y vaut g_0 ; il faudra en tenir compte dans la formulation du problème. En dehors de cela, il nous reste seulement à préciser quelle condition de régularité doit vérifier une fonction nulle sur Γ_0 pour appartenir à V. La condition (13) suggère de demander que son gradient ainsi qu'elle-même soient de carré intégrable. Or, c'est exactement la condition que dicte la physique. Dans les applications les plus usuelles, l'expression :

$$\int_\Omega \| \text{grad}\, u \|^2 dx,$$

appelée *intégrale de Dirichlet*, est à un coefficient près l'énergie du système étudié. C'est le cas en élasticité, en électrostatique, dans l'écoulement irrotationnel d'un liquide. L'ensemble des fonctions qui sont de carré intégrable, ainsi que leur gradient, a reçu le nom d'*espace de Sobolev* et on lui a attribué la notation $H^1(\Omega)$ (il y a des espaces de Sobolev plus généraux). V sera donc l'ensemble des fonctions qui appartiennent à $H^1(\Omega)$ et s'annulent sur Γ_0 (un théorème assure que cette dernière condition a bien un sens pour les fonctions appartenant à l'espace de Sobolev). On arrive finalement à la formulation suivante : « Trouver $u \in H^1(\Omega)$ telle que sa restriction à Γ_0 soit g_0 et que, pour tout $v \in V$, la relation (13) soit vérifiée. »

On notera que les conditions de Dirichlet et de Neumann ont des statuts très différents dans la formulation variationnelle : la première doit être imposée à u et aussi, par l'intermédiaire de la définition de V, à v, alors que la seconde est intégrée dans la relation (10), (11) ou (13) selon le cas.

La monotonie

Le fait d'admettre une formulation variationnelle du type (11) n'implique pas qu'une équation ou un système soit elliptique. Au demeurant, les méthodes d'étude liées à la formulation variationnelle admettent une extension au cas hyperbolique, c'est ce qu'on appelle la méthode des *inégalités d'énergie*. Ce qui caractérise l'ellipticité, c'est une propriété des fonctions F_i que nous allons aborder maintenant.

Les propriétés des solutions d'une équation aux dérivées partielles sont surtout déterminées par les termes contenant les dérivées de l'ordre le plus élevé (ici 2). Nous allons donc concentrer notre attention sur la dépendance en grad u des fonctions F_i de la formule (11). Nous supposerons que $F_0 = 0$, comme c'est d'ailleurs le cas dans l'équation de Poisson-Laplace et dans les équations de l'élasticité, entre autres. Enfin, nous examinons le cas d'une équation, le passage à un système n'implique pas ici d'idée nouvelle, seulement une complication du formalisme. Nous posons $F = (F_1, ..., F_n)$.

On dit qu'une fonction F est *monotone* si pour tout couple $(u, v) \in \mathbf{R}^n \times \mathbf{R}^n$:

$$(14) \quad (u - v) \,.\, (F(u) - (F(v)) \geqslant 0.$$

Cette terminologie est d'ailleurs assez malheureuse, puisque dans le cas d'une seule variable réelle elle amène à dire qu'une fonction croissante est monotone mais qu'une fonction décroissante ne l'est pas ! Il reste qu'elle est adoptée par tous les spécialistes. On dit que la fonction est strictement monotone si le seul cas d'égalité dans (14) est celui où $u = v$.

Si F est linéaire par rapport à grad u, on peut écrire :

$$F_i(\text{grad}\, u) = \sum_{j=1} a_{ij}(x, u)\frac{\partial u}{\partial x_j}\cdot$$

La condition de monotonie signifie alors que la partie symétrique de la matrice des a_{ij} est positive, et définie positive s'il y a monotonie stricte. On notera en particulier que dans le cas de l'équation de Poisson-Laplace c'est l'opérateur $-\Delta$ qui a la propriété de monotonie.

Montrons que si F est strictement monotone, deux solutions du problème variationnel ne peuvent différer que par l'addition d'une constante. Soient u_1 et u_2 ces deux solutions. Écrivons la relation variationnelle (11) pour chacune des deux avec la même fonction $v = u_1 - u_2$, puis retranchons l'une à l'autre les deux équations obtenues. Nous aboutissons à :

$$\int_\Omega (\text{grad}\, u_1 - \text{grad}\, u_2) \times [F(\text{grad}\, u_1) - F(\text{grad}\, u_2)]\, dx = 0.$$

Si $u_1 - u_2$ n'était pas constante, la fonction à intégrer dans le premier membre de cette équation serait positive et non nulle et, par suite, l'intégrale strictement positive.

Si F est strictement monotone, il suffit donc que l'espace V du problème variationnel ne contienne pas de constante non nulle pour qu'il y ait unicité ; ce sera le cas si on a imposé la valeur de la solution sur une partie de la frontière. Dans d'autres cas, c'est la présence d'un terme en F_0 strictement positif qui élimine la constante. Dans d'autres cas encore, tel celui du problème de Neumann, il existe bel et bien toute une famille de solutions différant deux à deux d'une constante.

Avec des conditions d'uniformité de la monotonie et des conditions de continuité assez faibles pour pouvoir être vérifiées dans la plupart des problèmes usuels, on démontre l'existence d'une solution (théorème de Minty-Browder). La démonstration, assez technique, se fait en deux étapes. La première consiste à démontrer l'existence dans le cas de la dimension finie. Elle s'appuie essentiellement sur un résultat de topologie algébrique (le « théorème des antipodes » de Borsuk). La seconde étape consiste à démontrer la convergence des approximations de Ritz-Galerkine.

Formulation variationnelle et calcul des variations

Dans de nombreux problèmes, parmi lesquels les plus fréquents dans les applications, la formulation variationnelle exprime que la solution u est point critique d'une fonctionnelle J sur l'espace V. Ainsi du problème mêlé pour l'équation de Poisson-Laplace (et comme cas particuliers des problèmes de Dirichlet et de Neumann). La fonctionnelle J est dans ce cas définie par la formule :

$$J(u) = \int_\Omega \left[\frac{1}{2}\|\text{grad}\, u\|^2 + uf\right] dx + \int_{\Gamma_1} u g_1\, dS.$$

Les équations linéaires du second ordre pour lesquelles une telle fonctionnelle peut se trouver sont celles qui s'écrivent :

$$\text{div}\,(A(x)\,.\,\text{grad}\, u) + c(x)\, u + f = 0,$$

où A est une matrice symétrique.

Lorsque cette fonctionnelle existe, on est donc ramené à un problème d'optimisation. La propriété de monotonie équivaut à la convexité de la fonctionnelle, qu'il s'agit donc de minimiser.

La fonctionnelle J a souvent une interprétation comme énergie potentielle du système. Sa convexité indique donc la stabilité de la configuration d'équilibre, elle lui est même équivalente dans le cas linéaire. Dans les formulations variationnelles, l'absence de fonctionnelle correspond souvent à une non-conservation de l'énergie, la condition de monotonie indiquant qu'il y a dissipation.

3. L'équation de la chaleur et le type parabolique

Si les équations hyperboliques décrivent l'évolution des phénomènes physiques réversibles, les phénomènes irréversibles relèvent du type parabolique dont le prototype est l'équation de la chaleur, dite aussi de Fourier :

$$\frac{\partial u}{\partial t} = \Delta_x u + f. \qquad (15)$$

Notons tout de suite qu'au contraire de l'équation des ondes cette équation est modifiée par le changement de t en $-t$.

Elle décrit la diffusion de la chaleur, mais aussi bien d'autres phénomènes de diffusion, en particulier celle d'un corps en solution.

Les problèmes bien posés typiques de l'équation de la chaleur, et des équations paraboliques en général, sont des *problèmes mixtes*. On donne un ouvert Ω de l'espace et on cherche une solution u sur $[0, \infty[\times \Omega$ qui vérifie une condition initiale : $u(0, x) = u_0(x)$, u_0 fonction donnée et, à chaque instant t, une condition sur la frontière, condition de Dirichlet, ou de Neumann ou mêlée, d'autres parfois. La différence avec le cas hyperbolique est à chercher dans le comportement vis-à-vis de la variable temps. D'abord on ne donne ici que la valeur initiale de u et pas celle de sa dérivée. Ensuite, et c'est le plus important, la solution n'existe en général que dans le futur, c'est-à-dire pour les valeurs positives de t. On retrouve là l'opposition réversibilité-irréversibilité. La coexistence de données initiales et de données à la frontière d'un ouvert d'espace n'est pas essentielle : on la trouve aussi dans certains problèmes hyperboliques.

Certaines propriétés de l'équation de la chaleur la rapprochent de l'équation de Laplace. Supposons pour le moment que $f = 0$ (c'est-à-dire qu'il n'y a ni source ni absorption de chaleur). Les solutions sont alors indéfiniment différentiables et, lorsqu'on fixe t, ce sont des fonctions analytiques de x. En particulier, la diffusion est instantanée dans ce sens que si dans un problème mixte la donnée initiale est nulle en dehors d'un voisinage d'un point, dès que t est strictement positif, il n'y a plus aucun point au voisinage duquel la solution reste nulle. Ce résultat peut aussi se déduire d'une version du principe du maximum adaptée à l'équation de la chaleur.

Nous allons établir l'équation de la chaleur par un raisonnement qui, convenablement modifié, s'étend aux autres phénomènes de diffusion. Nous prendrons pour u la densité de l'énergie interne, f est celle des sources d'énergie thermique. Nous négligeons les effets dus à la dilatation thermique, autrement nous n'arriverions pas à (15) mais à un système où figurerait cette équation et celles de l'élasticité, le tout modifié par un couplage de ces équations entre elles. Nous désigne-

rons par v le flux d'énergie thermique. Le bilan d'énergie dans un domaine U nous donne ainsi :

$$\frac{d}{dt}\int_U u\,dx = \int_U f\,dx - \int_{\partial U} v\,.\,n\,dS,$$

où ∂U désigne la frontière de U. Nous transformons une fois de plus l'intégrale de surface en intégrale de volume par une intégration par parties ; faisons passer la dérivation par rapport au temps sous le signe d'intégration (u peut-être supposée assez régulière pour que nous en ayons le droit) et obtenons :

$$\int_U \left[\frac{\partial u}{\partial t} + \operatorname{div}_x v - f\right] dx = 0.$$

Comme cette équation doit être vraie pour tout domaine U, on en déduit :

$$(16) \qquad \frac{\partial u}{\partial t} - \operatorname{div}_x v = f.$$

C'est l'*équation de continuité* dont la validité est extrêmement générale : elle exprime simplement une loi de conservation quelconque.

Il faut maintenant une relation entre u et v. C'est la loi de diffusion proprement dite qu'on prend de la forme :

$$(17) \qquad v = -K(x, u)\,.\,\operatorname{grad}_x u,$$

où K est a priori une matrice. Les physiciens démontrent qu'elle est symétrique. Elle est définie positive : cette propriété exprime que l'énergie thermique diffuse des régions les plus chaudes vers les plus froides et pas l'inverse. Dans un milieu isotrope, c'est simplement le produit par un nombre. Si le milieu est homogène elle ne dépend pas de x. Pour aboutir à l'équation (15) on suppose toutes ces propriétés vérifiées et on fait de plus l'hypothèse que K ne dépend pas non plus de u, ce qui est une approximation justifiée dans les situations les plus usuelles. Dans tous les cas, (16) et (17) se combinent en :

$$(18) \qquad \frac{\partial u}{\partial t} = \operatorname{div}(K(x, u)\,.\,\operatorname{grad} u) + f.$$

Si K est le produit par un nombre constant, un choix adéquat des unités donne la forme (15).

Dans le raisonnement qui précède, on peut remplacer l'énergie interne par la concentration d'une solution sans rien changer d'autre. Des raisonnements analogues s'appliquent aux fluides circulant dans un milieu poreux. Très simples dans le cas d'un liquide saturant les pores, les équations deviennent beaucoup plus compliquées dans le cas d'un gaz ou du mélange de deux fluides.

On notera que (18) peut s'écrire :

$$\frac{du}{dt} + A(u) = 0,$$

où A a la propriété de monotonie. C'est là la clef de propriétés d'existence et d'unicité pour les équations et systèmes paraboliques, sur lesquelles on reviendra dans la partie **C** ci-dessous - Équations non linéaires.

L'équation de la chaleur et les équations analogues ont un lien étroit avec la théorie des probabilités. On ne s'en étonnera pas puisque la relation (17) repose sur une théorie relevant de la physique statistique.

Pour nous borner à un aspect assez simple de cette question, considérons des solutions de l'équation :

$$(19) \qquad \frac{\partial u}{\partial t} = K\Delta u$$

vérifiant une condition de décroissance à l'infini sur $\mathbf{R}^n$ (par exemple intégrables). On

passe alors de la solution à l'instant t_1 à la solution à un instant ultérieur t_2 par la formule :

(20) $$u(t_2, .) = U_{t_2 - t_1} u(t_1, .),$$

où U_t est l'*opérateur d'évolution*, défini dans ce cas par la formule :

(21) $$U_t f(x) = \int_{\mathbf{R}^n} \frac{e^{-\|x-y\|^2/4Kt}}{(4\pi Kt)^{n/2}} f(y)\, dy.$$

Cet opérateur transforme les mesures de probabilités sur $\mathbf{R}^n$ en mesures de probabilités sur $\mathbf{R}^n$ et les gaussiennes en gaussiennes. C'est donc l'opérateur de transition d'un processus de Markov gaussien.

Réciproquement, on vérifie que tout processus de Markov gaussien à accroissements indépendants et stationnaires, commutant avec les déplacements de l'espace, est donné par les formules (20) et (21).

En plus des équations de diffusion, les systèmes paraboliques comprennent les équations de Navier-Stokes pour un fluide incompressible :

(22) $$\rho\left(\frac{\partial u}{\partial t} + \sum_j u_j \frac{\partial u}{\partial x_j}\right) - \mu \Delta u = -\operatorname{grad} p + f,$$

où u est la vitesse du liquide, p sa pression, ρ sa densité, μ son coefficient de viscosité et f la force extérieure. De plus, on impose à u d'être de divergence nulle : c'est l'équation de continuité.

4. Autres équations

Équations qui changent de type

L'équation de Tricomi :

$$x_2 \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0$$

est hyperbolique dans le demi-plan $x_2 < 0$, elliptique dans le demi-plan $x_2 > 0$.

En dehors de cela, le principal intérêt de l'équation de Tricomi est sa simplicité qui a permis d'en faire une étude assez détaillée. On rencontre un système présentant le même changement de type dans l'étude des écoulements stationnaires de fluides compressibles non visqueux. En admettant que l'équation d'état permet d'écrire la pression comme une fonction p de la densité, les équations du mouvement s'écrivent dans ce cas :

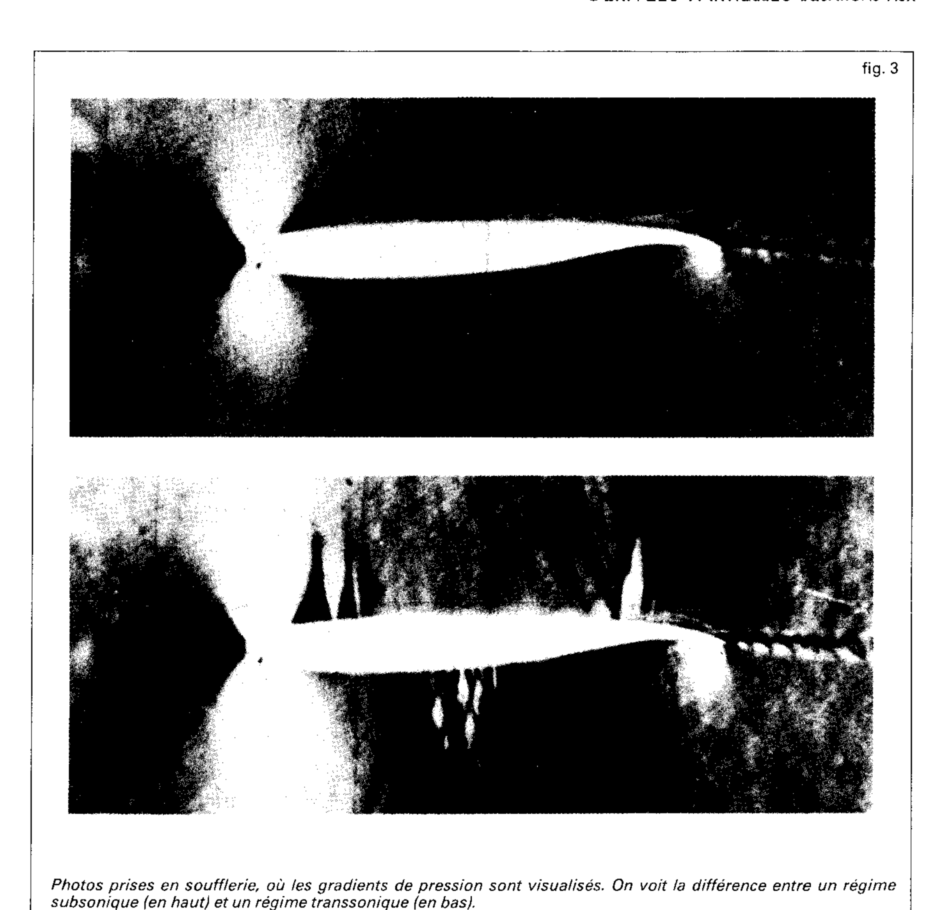
$$\rho \sum_j u_j \frac{\partial u}{\partial x_j} + p'(\rho) \operatorname{grad} \rho = 0,$$

$$\rho \operatorname{div} u + u \cdot \operatorname{grad} \rho = 0.$$

Les notations ont le même sens que dans l'équation (22) ; $p'(\rho)$ est le carré de la vitesse de propagation du son dans le fluide, on voit qu'elle dépend de x par l'intermédiaire de la densité. Dans la région subsonique, c'est-à-dire celle où $\|u\|^2 < p'(\rho)$, le système est elliptique, il est hyperbolique dans la région supersonique, c'est-à-dire celle où $\|u\|^2 > p'(\rho)$. Dans les équations et systèmes hyperboliques dont nous avons parlé jusqu'ici, la variable temps jouait un rôle privilégié ; ce rôle est tenu dans la région supersonique par le déplacement dans la direction de l'écoulement. On peut d'ailleurs dans ce cas entendre la propagation des singularités dans la région hyperbolique : c'est le « bang » de l'avion passant le mur du son.

Naturellement, c'est dans les régions comportant à la fois des parties supersoniques et des parties subsoniques que l'étude de l'écoulement est la plus délicate (régions transsoniques, fig. 3). L'étude de ces régions transsoniques connaît un regain d'intérêt car la hausse du prix du carburant a amené à renoncer, pour les avions de ligne du futur, aux vitesses supersoniques.

fig. 3



Photos prises en soufflerie, où les gradients de pression sont visualisés. On voit la différence entre un régime subsonique (en haut) et un régime transsonique (en bas).

L'équation de Schrödinger

L'équation de Schrödinger :

$$ih\frac{\partial u}{\partial t} = -\frac{h^2}{2m}\Delta u + \mathrm{V}u$$

décrit en physique quantique l'évolution de la fonction d'onde u d'une particule non relativiste de masse m soumise à un potentiel V. Elle n'est pas hyperbolique, ce qui semble mettre en défaut l'assertion selon laquelle la physique des phénomènes réversibles est décrite par des systèmes hyperboliques. Mais il ne faut pas oublier que l'équation de Schrödinger apparaît comme une approximation à faible vitesse de l'équation de Dirac qui, elle, est un système hyperbolique.

Malgré une certaine ressemblance formelle avec l'équation de la chaleur, elle n'est pas non plus parabolique ; le facteur i devant la dérivée par rapport au temps assure la réversibilité.

L'équation de Schrödinger n'appartient finalement pas à un type particulier d'équations. Malgré son importance physique, elle apparaît assez comme un cas particulier – au moins pour le moment.

Dans le même ordre d'idées, l'équation de Korteweg et de Vries et les autres équations « à solitons », dont les propriétés seront exposées dans la partie consacrée aux problèmes non linéaires, ont des caractères analogues : elles ne sont pas hyperboliques mais liées à des équations hyperboliques.

Équations générales

Occupons-nous d'abord des équations aux dérivées partielles linéaires à coefficients constants du second ordre à trois variables indépendantes. Par un changement de variables, on peut toujours se ramener à un des trois cas suivants :

a) la partie principale (homogène d'ordre deux) de l'équation est incomplète (c'est-à-dire qu'il n'y figure pas de dérivation par rapport à une des variables) ;

b) cette partie principale est l'opérateur de Laplace, l'équation est elliptique ;

c) la partie principale est l'opérateur des ondes (à deux dimensions d'espace), l'équation est hyperbolique.

Si on passe à quatre variables indépendantes, un quatrième cas se présente, celui de l'équation :

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}-\frac{\partial^2 u}{\partial x_3^2}-\frac{\partial^2 u}{\partial x_4^2}=f.$$

Cette équation est appelée ultrahyperbolique par allusion au fait qu'il y a un signe de plus que dans l'équation des ondes. Mais cette expression ne désigne pas un type comme l'hyperbolique ou le parabolique ; c'est au fond une désignation purement négative. Elle ne correspond ni à des propriétés de problèmes bien posés ou à des singularités de solutions, ni à une famille qu'on saurait caractériser sur la forme de l'équation, y compris pour les ordres supérieurs à deux.

Au demeurant, cette équation n'intervient pas dans des problèmes physiques.

En principe, la théorie de Sato, Kashiwara et Kawai doit permettre une classification des systèmes linéaires aux dérivées partielles. Malheureusement, elle est d'un caractère tellement abstrait et détourné qu'on n'a encore guère vu d'application à un cas particulier non classique.

MARTIN ZERNER

Bibliographie

Y. V. EGOROV, M. A. SHUBIN & R. V. GAMKRELIDZE, *Partial Differential Equations*, 2 vol., Springer-Verlag, New York, 1991-1992 / J. KEVORKIAN, *Partial differential Equations*, Chepman & Hall, 1990 / J. RAUCH, *Partial differential Equations*, Springer-Varlag, 1991.

B. Théorie linéaire

Il existe une théorie mathématique assez bien constituée des équations aux dérivées partielles linéaires, dont nous allons essayer de donner une idée. En contraste, les équations non linéaires présentent un foisonnement de problèmes et de méthodes dont peu sont générales. Sans que nous le précisions à chaque fois, certains des résultats que nous allons donner dans le cas linéaire se généralisent au non linéaire. Pourtant, même ceux-là font partie de la théorie linéaire, soit que la généralisation non linéaire soit limitée à des situations trop restrictives, soit qu'elle ne s'insère pas dans une théorie cohérente.

Pour pouvoir conserver les notations de la partie précédente pour les problèmes d'évolution, nous nous placerons sur un ouvert de $\mathbf{R}^{n+1}$ (à l'occasion $\mathbf{C}^{n+1}$) et nous noterons en général $y = (y_0, y_1,..., y_n)$ les

coordonnées. Pour $\alpha \in \mathbf{N}^{n+1}$, nous poserons :

$$|\alpha| = \sum_{j=0}^{n} \alpha_j$$

$$y^\alpha = y_0^{\alpha_0} y_1^{\alpha_1} \dots y_n^{\alpha_n}$$

$$\nabla^\alpha = \frac{\partial^{|\alpha|}}{\partial y_0^{\alpha_0} \partial y_1^{\alpha_1} \dots \partial y_n^{\alpha_n}}.$$

Un opérateur linéaire aux dérivées partielles (on dit plus brièvement *opérateur différentiel*) est défini par un polynôme à coefficients pouvant dépendre de y :

$$\mathrm{P}(y, \xi) = \sum_{|\alpha| \leqslant m} b_\alpha(y) \xi^\alpha ;$$

il agit selon la formule :

$$\mathrm{P}u(y) = \sum_{|\alpha| \leqslant m} b_\alpha \nabla^\alpha u(y).$$

m s'appelle l'*ordre* de P. L'opérateur P_m obtenu en ne gardant que les termes où la dérivation est d'ordre m exactement s'appelle la *partie princiale* de P. On notera ∇ le gradient :

$$\nabla_u = \sum_{i=0}^{n} \frac{\partial}{\partial y_i}$$

(cf. CALCUL INFINITÉSIMAL - Calcul à plusieurs variables).

À côté de ces notations, il nous arrivera d'utiliser des notations en (t, x) où $t \in \mathbf{R}$ (ou $\mathbf{C}$) et $x \in \mathbf{R}^n$ (ou $\mathbf{C}^n$) avec des conventions analogues pour les multi-indices, puissances et dérivations.

Nous ne nous priverons pas à l'occasion de noter les distributions comme des fonctions et les produits scalaires fonctions-distributions comme des intégrales (cf. DISTRIBUTIONS).

1. Le théorème de Cauchy-Kovalevskaïa

Supposons l'opérateur P de la forme :

$$(1) \quad \mathrm{P}u = \frac{\partial^m u}{\partial t^m} + \sum_{k=1}^{m} \mathrm{Q}_k(t, x, \nabla_x) \frac{\partial^{m-k} u}{\partial t^{m-k}},$$

où les Q_k sont des opérateurs différentiels d'ordre au plus k et où ∇_x désigne le gradient relativement à x.

Le problème de Cauchy s'énonce alors : « Trouver u vérifiant :

$$\begin{cases} (2) \quad \mathrm{P}u = f \\ (3) \quad \dfrac{\partial^k u}{\partial t^k}(0, x) = g_k(x), \quad k = 0, \dots, m-1, \end{cases}$$

où f et $g_0, g_1, \dots, g_{m-1}$ sont des fonctions données. »

Le *théorème de Cauchy-Kovalevskaïa* suppose que les coefficients de P ainsi que les données $f, g_0, \dots, g_{m-1}$ sont des fonctions analytiques (réelles ou complexes) de t et de x. Il affirme alors l'existence d'une solution analytique et une seule sur un voisinage de tout point $(0, x_0)$. Ce voisinage dépend de P et des domaines d'analyticité complexes des données.

Ce théorème s'applique aussi aux systèmes, pourvu qu'ils soient de la forme :

$$(4) \qquad \frac{\partial^m u}{\partial t^m} = \Phi,$$

où Φ est une fonction analytique de t, x, u et ses dérivées d'ordre total m au plus mais strictement plus petit que m en t. Il reste un des rares résultats très généraux de la théorie. Il a été publié par Cauchy en 1842 dans les *Comptes rendus de l'Académie des sciences*.

Le travail de Sofia Kovalevskaïa est paru en 1874 ; apparemment elle ne connaissait pas celui de Cauchy (et son jury non plus puisqu'il s'agissait d'une thèse !).

La démonstration d'unicité est simple et instructive. Si u est une solution analytique, elle possède un développement de Taylor en t :

$$u(t,x) = \sum_{k=0}^{\infty} \frac{1}{k!} g_k(x) t^k$$

où :

$$g_k(x) = \frac{\partial^k u}{\partial t^k}(0,x)$$

pour tout k positif cette fois. Les g_k sont donnés pour $k < m$. En faisant $t = 0$ dans l'équation aux dérivées partielles, on trouve :

$$g_m(x) = f - \sum_{k=0}^{m-1} Q_{m-k}(0, x, \nabla_x) g_k(x).$$

En dérivant l'équation (2) par rapport à t et en y faisant $t = 0$, on trouve des formules analogues donnant chaque g_k en fonc-
tion de ceux d'indices strictement plus petits.

La démonstration d'existence consiste essentiellement à démontrer la convergence de la série ainsi calculée. Elle repose sur une technique de majoration établie par Cauchy à cette occasion (méthode des séries majorantes).

La même démonstration s'applique d'ailleurs au système non linéaire (4) moyennant des complications légères.

Unicité de la solution distribution

Le théorème de Cauchy-Kovalevskaïa n'exclut pas l'existence de solutions non analytiques au problème de Cauchy. Cette lacune a été comblée, en 1901 par le *théorème de Holmgren* qui affirme l'unicité des solutions « classiques » (c'est-à-dire m fois différentiables). Très élégante, la démonstration de Holmgren est remarquable pour l'époque par la façon dont elle met en jeu des idées de l'analyse moderne : dualité et densité.

Le résultat de Holmgren a été étendu par Hörmander aux solutions distributions. Il est nécessaire dans ce nouveau cadre de reformuler le problème, puisque la restriction d'une distribution à l'hyperplan $t = 0$, qui intervient dans les données de Cauchy, n'a a priori pas de sens. Notons θ la fonction qui vaut 0 pour $t < 0$ et 1 pour $t \geqslant 0$, et δ la distribution de Dirac. La fonction u vérifie les conditions (2) et (3) si et seulement si on a l'équation entre distributions :

$$\mathbf{P}(\theta u) = \theta f + \sum_{k=0}^{m-k} \sum_{l=1}^{m-k} C_k^l \, \delta^{(l-1)}(t) Q_k(0, x, \nabla_x) g^{m-k-l}(x).$$

avec la convention $Q_0 = 1$

En généralisant cette formule, on est amené à poser le problème de Cauchy de la façon suivante : « Trouver une distribution u à support contenu dans le demi-espace $t \geqslant 0$ qui vérifie : $\mathrm{P}u = \mathrm{T}$, T distribution donnée à support contenu dans ce même demi-espace. »

L'unicité de la solution du problème de Cauchy devient ainsi une affirmation sur le support des distributions qui vérifient l'équation :

(5) $$\mathbf{P}u = 0.$$

Si le support d'une telle distribution est contenu dans le demi-espace fermé $t \geqslant 0$, il est aussi contenu dans le demi-espace ouvert $t > 0$.

Le problème de Cauchy en coordonnées générales : hypersurfaces caractéristiques

Dans certaines situations, on a besoin d'étudier un problème de Cauchy où les données, au lieu d'être portées par l'hyperplan $t = 0$, le sont par une autre hyper-

surface Σ. Il y a donc lieu de voir si on peut trouver des coordonnées (t, x) telles que :

a) l'opérateur P prend la forme (1) au produit près par une fonction non nulle (nous pourrons diviser le second membre par cette fonction) ; cela revient à dire que le coefficient de $(\partial^m u)/(\partial t^m)$ ne s'annule pas ;

b) l'équation de Σ devienne $t = 0$.

Supposons donc que Σ soit définie par une équation $S(y) = 0$, où S est une fonction analytique dont le gradient ne s'annule pas. Nous prenons pour nouvelles coordonnées $t = S(y)$ et des fonctions x_1, ..., x_n de façon que l'ensemble (t, x) fasse un système de coordonnées. Un calcul sans histoire montre que le coefficient de $(\partial^m u)/(\partial t^m)$ est :

$$P_m(y, \text{grad } S(y)). \quad (6)$$

Trois cas peuvent alors se présenter.

Le premier cas est dit non caractéristique : l'expression (6) est non nulle ; le théorème de Cauchy-Kovalevskaïa et celui de Holmgren s'appliquent au problème de Cauchy avec données portées par Σ.

Le deuxième cas est le cas caractéristique, c'est-à-dire que l'équation :

$$P_m(y, \text{grad } S(y)) = 0$$

est vérifiée. On dit que Σ est une *hypersurface caractéristique*. On peut démontrer dans ce cas que l'équation :

$$P(y, \nabla)u = 0 \quad (7)$$

a des solutions dont le support a Σ pour frontière, et des solutions dont les singularités sont portées par Σ.

Il reste des cas intermédiaires où l'expression (6) s'annule, mais pas identiquement. C'est le plus délicat. Il y a lieu à ce sujet de signaler les résultats de Leray sur l'uniformisation du problème de Cauchy : la solution se ramifie autour de la variété où l'expression (6) s'annule.

Un point important à retenir est que les hypersurfaces qui peuvent faire partie de la frontière du support d'une solution de l'équation (7) sont les caractéristiques. Un autre est la caractérisation des équations elliptiques : elles n'ont pas de caractéristiques réelles. Dans le cas elliptique, il n'y a pas d'hypersurfaces pour limiter le support des solutions de (7), et pour cause : on démontre qu'elles sont analytiques.

Les *limitations du théorème de Cauchy-Kovalevskaïa* ont été mises en lumière de façon particulièrement claire par Hadamard dans ses *Leçons sur le problème de Cauchy* (publiées à Yale en 1923 et à Paris en 1932). Elles portent sur trois points liés entre eux qui rendent le résultat inopérant dans les applications physiques :

- sa nature très locale ;
- l'hypothèse d'analyticité des données ; les situations physiques où l'analyticité est une propriété naturelle sont rares et en tout cas ce ne sont pas celles où se posent des problèmes de Cauchy ;
- conséquence des deux circonstances précédentes, une instabilité de la solution : si on modifie les données en leur ajoutant des fonctions analytiques si petites soient-elles, on perd tout contrôle de la solution, et même de son domaine d'existence, si on ne connaît pas le domaine d'analyticité complexe de la perturbation.

Ces limitations expliquent l'importance des équations hyperboliques définies comme celles où il y a encore existence pour le problème de Cauchy à données indéfiniment différentiables (ou à données distributions : si le passage de l'analytique au différentiable implique dans ce problème une différence essentielle, le passage des fonctions différentiables aux distributions est au contraire automatique pourvu

que les coefficients soient eux-mêmes indéfiniment différentiables).

2. Problèmes de régularité

On a déjà signalé que si P est un opérateur elliptique à coefficients analytiques et u une distribution vérifiant l'équation (2), u est analytique sur tout ouvert où f l'est. De plus, cette propriété caractérise les opérateurs elliptiques.

On dit que l'opérateur P est *hypoelliptique* si toute u vérifiant (2) est indéfiniment différentiable sur tout ouvert où le second membre f est indéfiniment différentiable.

Dans sa thèse, Hörmander a donné la caractérisation suivante des opérateurs hypoelliptiques à coefficients constants :

Pour tout α différent de 0 on a :

$$\lim_{\xi \to \infty} \nabla^{\alpha} P(\xi)/P(\xi) = 0.$$

La dérivée est évidemment prise par rapport à ξ : c'est la seule variable dont dépend P puisque les coefficients sont constants. L'intervention de ces dérivées est assez naturelle du fait qu'on cherche à localiser les propriétés de u en multipliant cette distribution par une fonction indéfiniment différentiable à support borné. On utilise alors la généralisation de la formule de Leibniz, valable pour tout opérateur différentiel linéaire :

$$P(\varphi u) = \sum_{\beta \leqslant m} \frac{\nabla_x^{\beta} \varphi}{\beta\,!} \nabla^{\beta} P u,$$

où $\beta\,!$ désigne le produit des factorielles de β_i.

Pour les opérateurs à coefficients variables (indéfiniment différentiables), on ne connaît que des conditions suffisantes d'hypoellipticité. Une de ces conditions s'exprime sur les opérateurs « à coefficients gelés », c'est-à-dire les opérateurs à coefficients constants obtenus, pour chaque point y, en remplaçant les coefficients variables b par leur valeur $b(y)$ désormais fixée. La condition est que chacun de ces opérateurs soit hypoelliptique et qu'ils aient tous le même domaine dans L^2. Une faiblesse de cette condition (obtenue à peu près simultanément par Hörmander et Malgrange) est qu'elle n'est pas conservée par les changements de coordonnées, comme le montre l'exemple de l'équation de la chaleur.

Par la suite, Hörmander a étudié les opérateurs de la forme :

$$P = \sum_{j=1}^{k} X_j^2 + X_0 + c,$$

où c est une fonction indéfiniment différentiable et $X_0, X_1, ..., X_k$ des opérateurs d'ordre un sans terme d'ordre zéro : chacun de ces opérateurs est donc défini par un champ de vecteurs (cf. la partie **A** ci-dessus - Sources et applications). Désignons par $[X_j, X_l]$ le commutant $X_k X_l - X_l X_k$; c'est encore un opérateur de la même nature et le champ de vecteurs qui lui correspond est le crochet des deux autres champs de vecteurs au sens de la géométrie différentielle. Nous noterons désormais de la même façon opérateurs du premier ordre et champs de vecteurs. Appelons encore Ξ le plus petit espace vectoriel stable par le crochet auquel $X_0, X_1, ..., X_k$ appartiennent et $r(y)$ la dimension de l'espace vectoriel formé par les valeurs au point y des champs appartenant à Ξ. Cet entier $r(y)$ prend son maximum m sur un ouvert non vide. Si m est strictement plus petit que la dimension $n + 1$ de l'espace, l'opérateur P n'est pas

hypoelliptique. En effet, d'après le théorème de Frobenius, on peut trouver un système de coordonnées locales dans lequel P ne contient pas de dérivations par rapport à certaines des variables. Hörmander démontre une réciproque partielle : si on a partout $r(y) = n + 1$, alors P est hypoelliptique.

3. Solutions élémentaires et paramétrix

On dit qu'une distribution de deux variables E est un *noyau élémentaire* de P si elle vérifie la relation :

$$P_y E(y, z) = \delta(y - z)$$

qui entraîne, du moins pour f à support compact, que la distribution :

$$u(y) = \int_{\mathbf{R}^{n+1}} E(y, z) f(z)\, dz$$

vérifie l'équation (2), d'où la précision noyau élémentaire *à droite* qu'il est prudent d'apporter, sauf, comme nous le verrons, dans le cas des équations à coefficients constants.

Nous avons déjà rencontré un tel noyau (cf. chap. 3 *L'équation de la chaleur et le type parabolique*, dans la partie A ci-dessus - Sources et applications, à propos du mouvement brownien). En effet les formules (20) et (21) de cette partie montrent que le noyau :

$$E(t_1, x_1, t_2, x_2) = (4\pi(t_1 - t_2))^{-n/2} \exp(-\|x_1 - x_2\|^2/4t)\, \theta(t_1 - t_2),$$

où $\theta(t) = 1$ pour t positif et 0 sinon, est un noyau élémentaire pour l'opérateur de la chaleur $\dfrac{\partial}{\partial t} - \Delta_x$.

Le plus ancien exemple de noyau élémentaire connu est sans aucun doute celui du potentiel coulombien $-1/4\pi\|y - z\|$, noyau élémentaire de l'opérateur de Laplace en dimension 3.

Opérateurs à coefficients constants et convolution

Un opérateur différentiel à coefficients constants est un opérateur de convolution puisqu'il commute avec les translations. Plus précisément :

$$Pu = P\delta * u.$$

Les noyaux élémentaires les plus commodes s'écrivent alors eux aussi comme noyaux de convolution $E(y - z)$, où E, qui ne dépend plus que d'une variable dans $\mathbf{R}^{n}{+1}$, est une *solution élémentaire*, c'est-à-dire qu'elle vérifie :

$$PE = \delta.$$

L'utilisation systématique de ce point de vue est un des traits caractéristiques du développement qu'a connu l'étude des équations aux dérivées partielles dans les années 1950 sous l'impulsion de la théorie des distributions. En particulier, Malgrange a démontré en 1953 que tout opérateur différentiel à coefficients constants non nul avait une solution élémentaire.

Solution élémentaire et hypoellipticité

Si P est un opérateur à coefficients constants hypoelliptique, ses solutions élémentaires doivent évidemment être indéfiniment différentiables en dehors de l'origine. Mais la réciproque est aussi vraie comme on va le voir. Il faut savoir que si T est une distribution indéfiniment différentiable en dehors de l'origine, alors, pour toute distribution f, le produit de convolution $T * f$ est indéfiniment différentiable sur tout ouvert où f l'est, pourvu qu'une au moins de ces deux distributions soit à support

compact. Supposons donc que P ait une solution élémentaire E indéfiniment différentiable en dehors de l'origine, et soit φ une fonction indéfiniment différentiable à support compact qui vaut 1 sur un voisinage de l'origine. Posons :

$$F = \varphi E.$$

Il est facile de s'assurer que :

$$PF = \delta + \psi,$$

où ψ est indéfiniment différentiable à support compact. On a donc :

$$F * f = F * P\delta * u = u + \psi * u,$$

la deuxième égalité à cause de l'associativité, et la commutativité du produit de convolution assurées du fait que u est le seul des trois facteurs à ne pas avoir un support compact. Comme $\psi * u$ est indéfiniment différentiable, on voit que u est indéfiniment différentiable sur tout ouvert où f est indéfiniment différentiable.

La distribution F utilisée dans cette démonstration est ce qu'on appelle une *paramétrix*. Le terme n'a pas de définition mathématique précise et universellement admise. Il signifie que PF est la distribution de Dirac plus « quelque chose de pas méchant », cette dernière expression désignant en général une fonction assez régulière.

Le résultat que nous venons de donner et sa démonstration s'étendent aux opérateurs à coefficients variables moyennant des complications techniques assez sérieuses.

Solution élémentaire et hyperbolicité

On se souvient que la formulation du problème de Cauchy en théorie des distributions amène à étudier l'équation aux dérivées partielles en supposant que second membre et solution ont leur support dans le « futur » (c'est-à-dire le demi-espace $t \geqslant 0$). Si P est hyberbolique, il faut en particulier (puisque le second membre peut être la distribution de Dirac) qu'il existe une solution élémentaire dont le support est contenu dans ledit futur (dans le cas des coefficients variables, un noyau élémentaire dont le support est contenu dans l'ensemble des couples (y, z) tels que y soit dans le futur de z).

Inversement, supposons que l'hyperplan $t = 0$ soit non caractéristique et qu'il existe une solution élémentaire E à support dans le futur. Supposons aussi, mais uniquement pour simplifier, que P soit à coefficients constants. Le théorème de Holmgren assure alors que l'hyperplan $t = 0$ n'a que l'origine en commun avec le support de E. Il y a plus : l'intersection de ce support avec tout hyperplan $t = C^{te}$ est un compact. Il en résulte que le produit de convolution $E * T$ est défini pour toute distribution à support dans le futur. Il résout le problème de Cauchy et par conséquent P est hyperbolique.

C'est la construction de solutions élémentaires qui a permis la démonstration d'existence dans le problème de Cauchy pour les opérateurs hyperboliques du second ordre à coefficients variables, quelques dizaines d'années avant que la théorie des distributions vienne fournir le cadre général dans lequel cette méthode s'insère aujourd'hui.

Solution élémentaire et répartition asymptotique de valeurs propres

Soit A un opérateur elliptique du second ordre ; pour étudier le problème de Dirichlet, restreignons-le aux fonctions qui s'annulent sur la frontière d'un ouvert borné Ω ; on obtient ainsi un opérateur auto-adjoint dans $L^2(\Omega)$ et cet opérateur est anticompact, c'est-à-dire que si un

nombre λ n'est pas valeur propre de A, alors $(A - \lambda I)^{-1}$ est un opérateur compact. Supposons-le inversible pour simplifier. Un noyau élémentaire, qui résout le problème de Dirichlet, est alors donné par la formule :

$$G(x,y) = \sum -\lambda_k^{-1}\, \psi_k(x)\, \psi_k(y),$$

où on a désigné par $-\lambda_k$ les valeurs propres de A (elles sont négatives, sauf peut-être un nombre fini d'entre elles) et par ψ_k une fonction propre normée associée à λ_k. Un tel noyau élémentaire qui résout un problème aux limites est en général appelé noyau de Green. Supposant les λ_k rangées par ordre croissant, nous allons nous intéresser à leur répartition asymptotique.

Pour avoir une idée de la difficulté de ce problème, on peut considérer le cas très simple où A est le laplacien et Ω un carré de côté 1. Les λ_k sont alors les nombres : $\pi^2(n_1^2 + n_2^2)$, n_1 et n_2 entiers, ce qui nous ramène à un problème célèbre en théorie des nombres, le nombre de points de coordonnées entières contenus dans un cercle de rayon r. Notons $N(r)$ le nombre de valeurs de k telles que :

$$\sqrt{\lambda_k} \leqslant r.$$

Il est facile de voir que :

$$N(r) = \frac{r^2}{4\pi} + o(r^2). \tag{8}$$

À la suite d'une série de travaux dont les premiers sont dus à Hermann Weyl, la partie principale de $N(r)$ est connue pour des problèmes elliptiques très généraux.

On se rendra compte que l'évaluation du terme en $o(r^2)$ dans la formule (8) est un problème très difficile si on sait que, dans ce simple cas particulier, de grands efforts ont été accomplis par les théoriciens des nombres pour obtenir son ordre de grandeur et que pourtant l'exposant de r n'y est pas encore exactement connu.

Un premier outil pour aborder ce problème s'obtient en évaluant la fonction :

$$\int e^{-rt}\, dN(r^2) = \sum e^{-\lambda_k t} = S(t).$$

C'est la trace (somme des valeurs propres) de l'opérateur U_t associé à l'équation parabolique :

$$\frac{\partial u}{\partial t} = Au, \tag{9}$$

en posant :

$$U_t g(.) = u(t, .),$$

où u vérifie (9) et $u(0, x) = g(x)$. La trace s'obtient en intégrant le noyau de U_t sur la diagonale. Le noyau de U_t, que nous noterons $U(t, x, y)$ s'écrit :

$$U(t,x,y) = \sum e^{-\lambda_k t}\, \psi_k(x)\, \psi_k(y),$$

de sorte que :

$$S(t) = \int U(t,x,x)\, dx.$$

Dans un travail paru en 1973, Colin de Verdière a mis ces idées en œuvre dans un contexte légèrement différent : A est l'opérateur de Laplace-Beltrami sur une variété riemannienne compacte X. Une construction par approximations successives de U lui permet de montrer que l'existence de géodésiques fermées et leur longueur influent sur le comportement asymptotique des valeurs propres.

Immédiatement après, Chazarain, d'une part, et Duistermat et Guillemin, d'autre part, ont obtenu des résultats analogues par une méthode légèrement différente dans laquelle la cause de l'inter-

vention des géodésiques fermées est plus apparente. Elle s'obtient en complétant N par antisymétrie ($N(-r) = -N(r)$) et en prenant la transformée de Fourier de sa dérivée :

$$\frac{1}{\sqrt{2\pi}}\int e^{i\lambda t}\,dN(\lambda) = \sum \cos\sqrt{\lambda_k}\,t = T(t),$$

équation purement symbolique entre distributions. La distribution T ainsi obtenue est associée à l'opérateur hyperbolique :

$$\frac{\partial^2}{\partial t^2} - A = P,$$

comme la fonction S l'était à un opérateur parabolique. En effet, si on définit l'opérateur V_t en posant :

$$V_t g(.) = u(t,.),$$

où u est la solution du problème de Cauchy :

$$\begin{cases} Pu = 0 \\ u(0,x) = g(x) \\ \dfrac{\partial u}{\partial t}(0,x) = 0, \end{cases}$$

on trouve que T est la « trace-distribution » de V_t. En fait, V_t n'est pas un opérateur à trace, mais, pour φ indéfiniment dérivable à support compact,

$$\int \varphi(t)\,V_t\,dt$$

en est un et sa trace n'est autre que :

$$\int T(t)\,\varphi(t)\,dt.$$

On remarquera que V_t s'obtient à partir d'une solution élémentaire de P en dérivant par rapport à t.

Si on se rappelle ce qui a été expliqué (cf. chap. 1 *L'équation des ondes et le type hyperbolique*, dans la partie A ci-dessus - Sources et applications, au sujet des bicaractéristiques), on ne s'étonnera pas des deux résultats suivants :

- les bicaractéristiques de P s'obtiennent en parcourant à une vitesse unité les géodésiques de X ;
- les singularités de V_t se propagent selon ces bicaractéristiques.

En particulier, si X possède une géodésique fermée de longueur L, on va voir revenir une singularité dans V_t avec une période L. Ce type de résultats a été étendu par la suite, en particulier au problème de Dirichlet. Le rôle des géodésiques fermées est alors joué par les lignes polygonales qui se referment par réflexion sur la frontière. On en déduit (pas directement !) que les singularités de T sont des points de la forme k L. Ce qui se traduit enfin sur le comportement asymptotique de N en vertu d'un des aspects de la dualité régularité locale-décroissance à l'infini dans la transformation de Fourier.

4. La transformation de Fourier et ses généralisations

Nous emploierons les notations suivantes pour la transformation de Fourier :

$$\mathcal{F}f(\xi) = \hat{f}(\xi) = \frac{1}{(2\pi)^{n/2}}\int e^{i<\xi,x>}f(x)\,dx,$$

où n est la dimension de l'espace (cf. DISTRIBUTIONS, chap. 4, et analyse HARMONIQUE, chap. 3).

Il en résulte que :

$$\mathcal{F}\frac{\partial f}{\partial x_j}(\xi) = i\xi_j\,\mathcal{F}f(\xi);$$

en d'autres termes, la transformation de Fourier transforme la dérivation partielle en produit par la variable correspondante, au facteur i près. Si P est un opérateur

différentiel *à coefficients constants* et u et f des distributions tempérées, l'équation aux dérivées partielles (2) équivaut à :

$$P(i\xi)\,\hat{u}(\xi) = \hat{f}\,(\xi).$$

Nous utiliserons le fait que la transformation de Fourier est inversible et a pour inverse $\overline{\mathcal{F}}$ définie par :

$$\overline{\mathcal{F}}g(x) = \frac{1}{(2\pi)^{n/2}}\int e^{-i<\xi,x>}\,g(\xi)\,d\xi$$

et le théorème de Parseval, étroitement lié au résultat précédent : $\mathcal{F}$ est une isométrie de $L^2(\mathbf{R}^n)$.

Une conséquence simple de ces résultats est que la transformation de Fourier et son inverse ont exactement les mêmes propriétés.

La dualité régularité locale-décroissance à l'infini

La transformée de Fourier d'une fonction intégrable est bornée. Si une fonction a des dérivées intégrables, sa transformée de Fourier décroît donc comme $1/\|\xi\|$, et si elle a des dérivées d'ordre k intégrables sa transformée de Fourier décroît à l'infini en $\|\xi\|^{-k}$. Inversement, si la transformée de Fourier $\hat{u}$ de u décroît à l'infini comme $\|\xi\|^{-k}$, u a des dérivées jusqu'à l'ordre $k-n-1$ qui sont continues et bornées.

Le décalage disparaît, grâce au théorème de Parseval, si on considère les fonctions de carré intégrable. Ainsi u appartient à l'espace de Sobolev $H^1(\mathbf{R}^n)$ (cf. chap. 2 *Le type elliptique*, dans la partie **A** ci-dessus - Sources et applications) si et seulement si sa transformée de Fourier est de carré intégrable pour la mesure $(1+\|\xi\|^2)\,d\xi$. Ce résultat fournit la définition la plus simple de l'espace de Sobolev d'indice réel quelconque : une fonction appartient à l'espace $H^s(\mathbf{R}^n)$ si sa transformée de Fourier est de carré intégrable pour la mesure $(1+\|\xi\|^2)^s\,d\xi$. L'introduction d'indices non entiers est nécessaire surtout du fait que, pour $s > 1/2$, on sait définir la restriction à un hyperplan d'une fonction appartenant à H^s, et cette restriction appartient à $H^{s-1/2}$; de plus, toute fonction appartenant à $H^{s-1/2}$ de l'hyperplan est restriction d'une fonction appartenant à H^s de l'espace ambiant.

Un cas extrême de décroissance à l'infini est donné par un support compact. Il lui correspond du côté Fourier une propriété d'analyticité : une distribution est à support compact si et seulement si sa transformée de Fourier se prolonge en une fonction analytique sur $\mathbf{C}^n$ tout entier à croissance exponentielle à l'infini (théorème de Paley Wiener généralisé). Il est bon de remarquer ici que la croissance exponentielle à l'infini de l'extension à $\mathbf{C}^n$ apparaît en même temps comme une propriété locale de la fonction puisqu'on peut la caractériser sur la suite des dérivées en un point via le développement de Taylor.

Propriétés des solutions élémentaires

Tout polynôme non nul possède un inverse multiplicatif qui est une distribution tempérée. Par transformation de Fourier, cela revient à dire que tout opérateur différentiel à coefficients constants possède une solution élémentaire tempérée.

Voyons d'abord comment ce résultat permet de démontrer l'hypoellipticité des opérateurs elliptiques. Soit P un tel opérateur. L'ellipticité signifie que la partie principale $P_m(\xi)$ n'a pas de zéro réel non nul et, par homogénéité, il en est de même de $P_m(i\xi)$. Soit E une solution élémentaire tempérée. On a pour tout α :

$$\nabla^{\alpha}\hat{E} = Q_{\alpha}/P^{|\alpha|+1}$$

où Q_α est un polynôme de degré $|\alpha|(m-1)$. Par conséquent, $\nabla^\alpha \hat{E}$ décroît à l'infini comme $1/\|\xi\|^{m+|\alpha|}$. En particulier, $\Delta^k \hat{E}$ décroît comme $\|\xi\|^{(m+2k)}$, ce qui montre, en revenant « côté x », que $\|x\|^{2k}E$ est $m+2k-\mathrm{n}-1$ fois continûment dérivable et, par division, que E est $m+2k-\mathrm{n}-1$ fois dérivable en dehors de l'origine. Mais k peut être choisi aussi grand qu'on veut, d'où l'indéfinie différentiabilité de E en dehors de l'origine.

Ce raisonnement s'adapte d'ailleurs au cas général des opérateurs hypoelliptiques à coefficients constants, à condition de savoir que, si $P(\nabla)$ est hypoelliptique, $P(i\xi)$ croît à l'infini comme une puissance strictement positive de $\|\xi\|$. Cette propriété se démontre grâce à une combinaison du développement de Puiseux et du théorème de Tarski-Seidenberg (l'image par une application polynomiale d'un ensemble défini par des équations et inéquations polynomiales est un ensemble de même nature).

Passons aux opérateurs hyperboliques. Il s'agit ici de savoir s'il y a une solution élémentaire dont le support est contenu dans le demi-espace $t \geqslant 0$ et qui coupe tout hyperplan $t = C^{te}$ selon un compact. Le résultat essentiel est une généralisation du théorème de Paley Wiener, le théorème de Plancherel. Sous sa forme la plus simple, il dit qu'une distribution tempérée d'une variable est à support contenu dans la demi-droite $x \geqslant a$ si et seulement si sa transformée de Fourier est limite, pour η tendant vers 0, d'une fonction F analytique dans le demi-plan complexe $\eta > 0$ qui vérifie une inégalité :

$$F(\xi + i\eta) \leqslant C|\xi|^k e^{-a\eta}.$$

À plusieurs variables, le théorème exprime une dualité entre distributions à support dans un convexe et fonctions analytiques sur $\mathbf{R}^n + i\Gamma$ vérifiant une condition de croissance, où Γ est un cône convexe.

On conçoit donc que P sera hyperbolique si $P(i\tau, i\xi)$ est non nul sur un ensemble $\mathbf{R}^{n+1} + i\Gamma$. Toutefois, sous cette forme, la condition n'est pas nécessaire parce que la solution élémentaire à support dans le futur peut ne pas être tempérée. Une condition nécessaire et suffisante d'hyperbolicité est que l'hyperplan $t = 0$ soit non caractéristique et qu'il existe τ_0 tel que :

$$P(\tau' + i\tau'', i\xi) \neq 0 \quad \text{pour} \quad \tau' > \tau_0.$$

Mais cette condition est équivalente à la condition apparemment plus forte : il existe τ_0 et un cône convexe ouvert Γ tels que :

$$P(\tau' + i\tau'', \eta + i\xi) \neq 0 \quad \text{pour} \quad \tau' > \tau_0 \quad \text{et} \quad \eta \in \Gamma.$$

Une condition nécessaire plus simple est que, pour tout ξ réel, le polynôme en τ $P_m(\tau, \xi)$ ait toutes ses racines réelles ; une condition suffisante est que, de plus, ces racines soient simples pour $\xi \neq 0$. Ces conditions portant sur la partie principale se généralisent aux opérateurs à coefficients variables.

Opérateurs pseudo-différentiels et opérateurs intégraux de Fourier

On voit qu'on dispose d'outils puissants pour l'étude des équations aux dérivées partielles à coefficients constants. Le passage aux coefficients variables a souvent consisté à se ramener aux coefficients constants. L'idée est que, les coefficients étant continus, localement l'équation est assez proche d'une équation à coefficients

constants, ce qui permet soit des majorations, soit des approximations successives.

À partir des années 1960, on a développé des outils permettant de travailler plus systématiquement sur les équations à coefficients variables. Grâce aux propriétés de la transformation de Fourier, on peut écrire :

$$(10)\quad Pu(x) = (2\pi)^{-n} \iint_{\mathbf{R}^n \times \mathbf{R}^n} e^{i<\xi, x-y>} P(x, i\xi)\, u(y)\, d\xi\, dy.$$

On généralise cette formule en n'y supposant plus que P soit un polynôme en ξ. On dit qu'on a affaire à un *opérateur pseudo-différentiel* d'ordre m s'il vérifie les inégalités :

$$\sup_{x \in K} \left| \nabla_x^\alpha \nabla_\xi^\beta P(x, \xi) \right| \leqslant C_{K,\alpha,\beta} \|\xi\|^{m-|\beta|},$$

pour tout compact K et tout ξ différent de 0. Le cas effectivement utilisé est celui où P est somme d'une fonction positivement homogène de degré m en ξ et d'un deuxième opérateur pseudo-différentiel d'ordre strictement plus petit que m.

Les opérateurs pseudo-différentiels permettent de plonger les opérateurs différentiels dans une algèbre où se trouvent aussi leurs paramétrix. Ils jouent ainsi le rôle qui est pour les opérateurs à coefficients constants, celui de la convolution. Ce cadre est très bien adapté au calcul de paramétrix, du moins dans le cas elliptique. Dans les autres cas, on en tire parti en passant au point de vue microlocal.

L'étape suivante est le passage aux *opérateurs intégraux de Fourier*. La généralisation consiste à remplacer dans la formule (10) la phase $<\xi, x-y>$ par une fonction plus générale.

Un cas particulier de cette méthode avait été employé par les physiciens en optique ondulatoire, c'est le développement en ondes sphériques. Pour une onde monochromatique :

$$u(t,x) = e^{i\omega t} v(x),$$

l'équation des ondes devient l'équation de Helmholtz :

$$(11)\qquad \Delta v + k^2 v = 0,$$

où $k = \omega/c$. Cette équation admet la solution élémentaire :

$$S(x) = -\frac{1}{4\pi\|x\|} e^{ik\|x\|},$$

d'où des solutions de la forme :

$$\frac{-1}{4\pi} \int \frac{e^{ik\|x-y\|}}{\|x-y\|} f(y)\, dy$$

(il s'agit soit de l'équation avec second membre correspondante, soit plus souvent de fonctions qui vérifient (11) en dehors du support de f). Une intégration par rapport à k permet de passer à d'autres solutions de l'équation des ondes.

MARTIN ZERNER

Bibliographie

H. BREZIS, *Analyse fonctionnelle*, Masson, Paris, 1983 / J. CHAZARAIN & A. PIRIOU, *Introduction à la théorie des équations aux dérivées partielles linéaires*, Gauthier-Villars, Paris, 1981 / J. HADAMARD, *La Théorie des équations aux dérivées partielles*, Éd. scientifiques, Pékin, 1964 / L. HÖRMANDER, *Analysis of Linear Partial Differential Equations*, t. I et II, Springer, 1985 / M. A. PINSKY, *Partial differential Equations and Boundary Value Problems*, McGraw-Hill, New York, 2e éd., 1991 / F. TREVES, *Pseudodifferential and Applications*, American Mathematical Society, Providence (R.I.), 1985.

C. Équations non linéaires

L'étude des équations aux dérivées partielles non linéaires se trouve à l'*interface* de nombreux problèmes scientifiques. En effet, la plupart des phénomènes de la physique ou des sciences de l'ingénieur sont non linéaires et une modélisation par des équations linéaires risque, dans certains cas, d'effacer des *événements* que les équations linéaires ne peuvent pas prendre en compte. Inversement, on peut dire que c'est l'existence de ces phénomènes nouveaux - apparition de chocs ou de singularités, comportement asymptotique profondément différent de celui des problèmes linéaires - qui rend la théorie difficile et qui conduit à faire appel à un arsenal mathématique très vaste. L'interaction avec le reste de la mathématique se fait aussi en sens inverse, car un certain nombre de problèmes *abstraits* se traitent à l'aide d'équations aux dérivées partielles non linéaires. Les liens avec l'analyse numérique sont continuels, et s'effectuent dans les deux sens. D'une part, on utilise l'analyse des équations aux dérivées partielles non linéaires pour construire des algorithmes numériques utilisés de plus en plus systématiquement. D'autre part, on se sert de l'ordinateur comme outil d'investigation. On effectue des calculs approchés concernant des phénomènes sur lesquels on ne possède que très peu d'information et, de ces calculs approchés, on déduit des conjectures que l'on s'efforcera par la suite de démontrer. Cette démarche, pressentie par John von Neumann, s'est révélée particulièrement féconde.

Bien entendu, un certain nombre de questions propres aux problèmes linéaires peuvent se généraliser aux problèmes non linéaires si, d'une part, les perturbations dues aux non-linéarités sont petites, et si, d'autre part, la structure des problèmes linéarisés correspondants introduit assez de régularité. Il en est ainsi des théorèmes d'existence des solutions de systèmes elliptiques ou paraboliques non linéaires et du comportement asymptotique de solutions d'équations du type :

$$\frac{\partial^2 u}{\partial t^2} - \Delta u + \mathrm{F}(u) = 0,$$

lorsque $\mathrm{F}(u)$ est une non-linéarité d'ordre assez élevé pour introduire un terme négligeable pour u petit. Des problèmes de ce type interviennent par exemple en théorie de la diffusion non linéaire. Plutôt que de développer un tel point de vue, nous allons décrire des problèmes où la non-linéarité joue un rôle dominant. Dans ces exemples, nous dégagerons deux idées. La première idée est que les solutions sont en général peu régulières et donc que les solutions ne pourront avoir un sens qu'en utilisant la théorie des distributions. Encore plus que dans le cadre linéaire, cette théorie s'impose dans le cadre non linéaire. Cette situation introduit une difficulté supplémentaire pour la construction de solutions ou le passage à la limite dans les méthodes approchées. En effet, comme il s'agit d'un problème non linéaire, on sera conduit à étudier la limite d'expressions de la forme $\mathrm{F}(u_n)$, où F est non linéaire. En général, il sera facile de prouver que (u_n) converge vers une fonction u et $\mathrm{F}(u_n)$ vers une fonction G (au sens des distributions), mais il sera difficile de prouver que $\mathrm{F}(u) = \mathrm{G}$. Par exemple, pour n tendant vers l'infini, la fonction $u_n(x) = \sin nx$ converge vers 0 au sens des distributions, tandis que la fonction :

$$\mathrm{F}(u_n(x)) = (u_n(x))^2 = \sin^2 nx = \frac{1}{2}(1 - \cos 2nx)$$

converge vers 1/2, toujours au sens des distributions.

Bien entendu, cette pathologie s'explique par le fait que, pour n tendant vers l'infini, la fonction sin nx oscille de plus en plus vite. On sera donc amené à montrer que, dans un sens convenable, les oscillations des solutions approchées ne sont pas trop grandes ; mais, justement, ce point est en contradiction avec l'existence de solutions singulières, et une analyse très fine est alors nécessaire.

La seconde idée est de faire appel, beaucoup plus que dans le cas linéaire, à la comparaison avec les équations différentielles ordinaires. Par exemple, on peut montrer qu'une onde de choc se forme en comparant la dérivée de la vitesse à la solution de l'équation différentielle ordinaire :

$$y' = y^2,$$

donnée par $y(t) = 1/(1 - ty_0)$, qui devient infinie en un temps fini. De tels phénomènes apparaissent aussi dans la description de la focalisation de rayons lasers. On peut montrer de plus que, pour des temps grands, les solutions de certaines équations (Navier-Stokes par exemple) restent proches de problèmes en dimension finie. Inversement, des problèmes non linéaires à très peu de degrés de liberté peuvent exhiber beaucoup de pathologie.

Ainsi, il n'existe pas de théorie générale du non-linéaire et cet article est une collection d'exemples significatifs. La non-linéarité intervient dans beaucoup d'autres problèmes, en particulier dans les équations elliptiques qui décrivent les surfaces minimales, dans les équations de Monge-Ampère et dans les équations de Yang-Mills de la théorie quantique des champs.

1. Les systèmes hyperboliques non linéaires

On se propose de considérer des systèmes de la forme :

$$\text{(1)} \qquad \frac{\partial u}{\partial t} + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \mathrm{F}_i(u) = 0,$$

où u est un vecteur à m composantes et F_i une fonction régulière de $\mathbf{R}^m$ dans $\mathbf{R}^m$. Son gradient (par rapport à u) est donc une matrice $\mathrm{A}_i'(u)$, et on dira que le système (1) est non linéaire hyperbolique si les A_i sont des fonctions non linéaires de u et si les valeurs propres de la matrice :

$$\sum_{i=1}^{n} \mathrm{F}_i'(u)\xi_i$$

sont toutes réelles pour tout vecteur $\xi = (\xi_1, \xi_2, ..., \xi_m) \in \mathbf{R}^m$.

De tels systèmes se rencontrent dans de nombreux domaines de la physique (mécanique des fluides, magnétohydrodynamique, combustion, etc.). Ils correspondent à des problèmes physiques célèbres, comme le calcul de la traînée et de la portance d'une aile d'avion, ou la propagation d'une onde de choc.

Pour comprendre la difficulté du problème, on peut considérer un modèle « abstrait » qui décrit la distribution des vitesses d'un fluide monodimensionnel sans force extérieure. Le mouvement des particules est donné par l'équation différentielle ordinaire :

$$\text{(2)} \qquad \dot{x}(t) = u(x(t), t),$$

et la relation fondamentale de la mécanique conduit à écrire :

$$\text{(3)} \quad 0 = \ddot{x}(t) = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\dot{x}(t) = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}u = \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right).$$

L'équation ainsi obtenue est dite *équation de Burger.* On déduit de (3) que l'on a :

$$\frac{d}{dt}u(x(t), t) = 0,$$

et donc que u est constant le long des solutions de (2), ce qui implique ensuite que les courbes définies par (2) sont en fait des droites. Cela permet de construire la solution de l'équation de Burger :

$$(4) \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0, \quad u(x, 0) = \varphi(x),$$

pour une donnée initiale régulière φ, et pendant un temps petit, en inversant l'équation :

$$x = t\,\varphi(\xi_x) + \xi_x,$$

et posant $u(x, t) = \varphi(\xi_x)$. Un tel procédé n'est possible que tant que l'équation (4) est résoluble. Or, l'équation (4) cesse d'être résoluble dès que deux droites caractéristiques se rencontrent, ce qui correspond à un choc entre les molécules de fluides et empêche que la solution reste continue. On est donc conduit à chercher une solution discontinue de (3) après le choc, c'est-à-dire une fonction u-mesurable et bornée qui vérifie (4) au sens des distributions. En particulier, si elle est discontinue le long d'une courbe $x = s(t)$, dite courbe de choc, elle devra vérifier la relation de saut :

$$s'(t) = \frac{1}{2}(u^+ + u^-),$$

où u^+ et u^- désignent les vitesses du fluide avant et après le choc. Une telle relation, qui est contenue dans la formulation (4), au sens des distributions, est dite relation de Rankine-Hugoniot. En fait, un choc correspond à une perte d'information et, sur des exemples simples, on voit que la relation (4) ne suffit pas à assurer l'unicité de la solution. Il convient d'ajouter une condition supplémentaire qui signifie que les caractéristiques rentrent dans le choc. Une telle condition est dite condition d'entropie et elle est, dans cet exemple, équivalente à la condition $u^+ < u^-$ qui signifie que, dans un choc, la vitesse diminue. On est donc conduit à résoudre l'équation (4) dans le cadre des fonctions qui admettent des discontinuités et qui vérifient de plus la *relation d'entropie.* La vérification de cette relation d'entropie introduit une difficulté car elle fait intervenir les courbes de choc, qui sont elles-mêmes des inconnues du problème. Pour simplifier, on introduit une fonction $\eta(u)$, strictement convexe, quelconque et on note $g(u)$ la primitive de la fonction $\eta'(u) \,.\, u$; on remarque que là où u est une solution régulière, elle vérifie la relation :

$$(5) \qquad \frac{\partial}{\partial t}\eta(u) + \frac{\partial}{\partial x}g(u) = 0.$$

Cela cesse d'être vrai là où il y a des discontinuités, car la condition de saut pour le premier membre de (5) est en général différente de la relation de Rankine-Hugoniot. Ainsi :

$$\frac{\partial}{\partial t}\eta(u) + \frac{\partial}{\partial x}g(u)$$

est une distribution dont le support est porté par l'ensemble des points singuliers de u. Il est ensuite facile de voir que la relation $u^- > u^+$ est équivalente au fait que cette distribution est négative. Une solution faible entropique est donc une fonction qui vérifie au sens des distributions les équation et inéquation :

$$(6) \quad \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0, \quad \frac{\partial \eta}{\partial t}(u) + \frac{\partial g}{\partial x}(u) \leqslant 0.$$

Dans le cas scalaire, on peut généraliser ces notions à un système à n variables d'espace :

$$(7) \qquad \frac{\partial u}{\partial t} + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(u) = 0.$$

À toute fonction $\eta(u)$, on associe alors une fonction $G(u)$ définie par :

$$(8) \qquad G(u) = (g_1, g_2, ..., g_n) ; \quad g_i'(u) = \eta'(u) F_i'(u),$$

et on démontre qu'il existe une unique fonction u, solution au sens des distributions dans $\mathbf{R}_x^n \times [0, \infty]$ du problème :

$$(9) \qquad \frac{\partial u}{\partial t} + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(u) = 0,$$

$$\frac{\partial \eta}{\partial t}(u) + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} g_i(u) \leqslant 0,$$

$$u(x, 0) = \varphi(x),$$

où φ désigne la donnée initiale, et η une fonction quelconque strictement convexe. Un des outils essentiels ici est l'utilisation des espaces de fonctions à variation bornée qui, en gros, sont des espaces de fonctions dont les dérivées, au sens des distributions, sont des mesures. Ces espaces sont assez vastes pour contenir des fonctions discontinues mais suffisamment réguliers pour satisfaire à des théorèmes de convergence adaptés aux problèmes non linéaires. Par exemple, la suite de fonctions déjà citée $u_n(x) = \sin nx$ n'est pas uniformément à variation bornée. Une idée de base pour la construction de la solution est d'introduire l'équation perturbée :

$$(10) \qquad \frac{\partial u_\varepsilon}{\partial t} + \sum \frac{\partial}{\partial x_i} F_i(u_\varepsilon) = \varepsilon \Delta u_\varepsilon, \quad u(x, 0) = \varphi(x).$$

C'est une équation parabolique non linéaire dont la résolution est beaucoup plus facile car la présence du laplacien empêche la formation de chocs. Lorsque ε tend vers zéro, u_ε converge vers une solution de :

$$\frac{\partial u}{\partial t} + \sum \frac{\partial}{\partial x_i} F_i(u) = 0.$$

De plus, si on multiplie par $\eta'(u_\varepsilon)$, on obtient :

$$(11) \qquad \frac{\partial}{\partial t} \eta(u_\varepsilon) + \sum \frac{\partial}{\partial x_i} g_i(u_\varepsilon) = \varepsilon \Delta \eta(u_\varepsilon) - \varepsilon \eta''(u_\varepsilon) |\nabla u_\varepsilon|^2.$$

Comme $\eta''(u_\varepsilon)$ est positif, on en déduit la relation :

$$\frac{\partial}{\partial t} \eta(u_\varepsilon) + \sum \frac{\partial}{\partial x_i} g_i(u_\varepsilon) \leqslant \varepsilon \Delta \eta(u_\varepsilon).$$

Il est facile de montrer que u_ε est borné et donc que :

$$\varepsilon \Delta \eta(u_\varepsilon) = \Delta(\varepsilon \eta(u_\varepsilon))$$

converge vers zéro au sens des distributions. On obtient ainsi l'inégalité d'entropie.

Dans la plupart des problèmes physiques, u n'est pas un scalaire mais un vecteur, car le système est décrit par plusieurs paramètres ; en particulier, en mécanique des fluides, par la densité, les composantes de la vitesse et la température. On obtient donc des systèmes de la forme :

$$(12) \qquad \frac{\partial u}{\partial t} + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} F_i(u) = 0.$$

Comme dans le cas scalaire, ces systèmes génèrent des ondes de choc, donc il est indispensable de chercher des solutions faibles. On est amené à généraliser la relation de saut (cf. *supra*) et la condition d'entropie. Pour l'entropie, une condition supplémentaire apparaît car $\eta(u)$ est maintenant une fonction de m variables réelles.

On peut multiplier à gauche l'équation (12) par le vecteur $\nabla\eta(u)$, mais il n'existe pas en général de fonction $g_i(u)$ de m variables qui vérifie la relation :

$$(13)\quad \nabla g_i(u) = \nabla\eta(u)\,\nabla F_i(u) = (K_1(u), K_2(u), ..., K_m(u)),$$

car le vecteur figurant à droite de (13) n'est en général pas un vecteur gradient ; il faut (et, dans les exemples que l'on considère en mécanique des fluides, il suffit) que l'on ait les relations :

$$\frac{\partial K_i(u)}{\partial u_j} = \frac{\partial K_j(u)}{\partial u_i}.$$

Un calcul simple montre que ces relations sont équivalentes à la relation matricielle :

$$(14)\quad \nabla^2\eta(u)\,.\,\nabla F_i(u) = (\nabla F_i(u))^*\,.\,\nabla^2\eta(u),$$

ou, ce qui revient au même, à la symétrie des matrices $\nabla^2\eta(u)\,.\,\nabla A_i(u)$. Il se trouve que la plupart des systèmes hyperboliques possèdent une fonction $\eta(u)$ qui vérifie les propriétés de commutation (14). Cette fonction coïncide, à un éventuel changement de signe près, avec la fonction d'entropie introduite en thermodynamique. Ce fait justifie a posteriori le nom de fonction d'entropie pour $\eta(u)$ et de flux d'entropie pour $g(u) = (g_1(u), g_2(u), ..., g_n(u))$. Cette notion a deux conséquences importantes. D'une part, elle conduit à caractériser les solutions des systèmes par les équation et inéquation suivantes :

$$(15)\quad \frac{\partial u}{\partial t} + \sum \frac{\partial}{\partial x_i} F_i(u) = 0\,;$$

$$\frac{\partial}{\partial t}\eta(u) + \sum \frac{\partial}{\partial x_i} g_i(u) \leqslant 0.$$

satisfaites *au sens des distributions*. D'autre part, dans la phase de régularité (par exemple pour un temps petit), avec des données initiales régulières, le système (12) est équivalent à :

$$(16)\quad \nabla^2\eta(u)\frac{\partial u}{\partial t} + \sum \nabla^2\eta(u)\,F_i'(u)\frac{\partial u}{\partial x_i} = 0.$$

Si on utilise ce système pour étudier des perturbations autour d'un état stationnaire constant $\bar{u}$, on est conduit aux équations :

$$(17)\quad \nabla^2\eta(\bar{u})\frac{\partial u}{\partial t} + \sum \nabla^2\eta(\bar{u})\,F_i'(\bar{u})\frac{\partial u}{\partial x_i} = 0,$$

qui constituent un système symétrique hyperbolique.

Ces considérations justifient l'importance pratique des systèmes symétriques, dont les applications s'étendent bien au-delà des équations de Maxwell et de Dirac.

Bien que les systèmes hyperboliques non linéaires aient suscité depuis longtemps l'intérêt des mathématiciens (par exemple, en 1930, Courant et Friedrichs étudiaient les problèmes hyperboliques posés par le vol supersonique), on ne dispose pas de théorèmes permettant de prouver dans un cadre assez général l'existence et l'unicité de la solution d'un tel système. Pour u vecteur, le seul théorème actuellement disponible est relatif à une seule variable d'espace. Une démonstration a été donnée par Glimm en 1966 sous des hypothèses assez générales ; à part un traitement particulier du problème de l'élasticité non linéaire, dû à Di Perna, elle n'a pas été améliorée depuis. Le modèle monodimensionnel est cependant très étudié car il correspond par exemple à la propagation d'une onde de choc dans un piston de moteur à explosion. Il permet aussi de considérer des problèmes possédant une symétrie sphérique, comme le problème de l'explosion de la poudre autour des masses critiques d'une bombe atomique.

2. Les équations de Navier-Stokes

Le chapitre précédent était consacré aux systèmes hyperboliques non linéaires, domaine où la différence entre le comportement des problèmes linéaires et les comportements des problèmes non linéaires apparaît de manière très évidente. Mais ces systèmes présentent les inconvénients suivants :

Il n'existe que des résultats partiels et la plupart des questions restent largement ouvertes.

Les applications concernent surtout la mécanique des fluides compressibles. Les lois de conservation classiques donnent alors un système d'équations qu'il faut compléter par une *loi d'état*, par exemple $p = \mathrm{R}\rho\mathrm{T}$ (pour les gaz parfaits) ; cette loi dépend du modèle considéré et peut être obtenue soit par des arguments physiques, soit (sans qu'aucune justification mathématique ne soit actuellement disponible) à partir de l'équation de Boltzmann (calcul des coefficients de transport par la méthode de Chapman-Enskog). On a donc un système qui est toujours compliqué et particulier.

Pour les raisons qui précèdent, l'intérêt s'est porté sur les équations d'Euler ou de Navier-Stokes, qui s'obtiennent en supposant le fluide incompressible mais en considérant éventuellement des termes de viscosité.

Les équations de Navier-Stokes et d'Euler présentent les propriétés suivantes. Elles sont (lorsque le problème est posé dans $\mathbf{R}^2$ ou dans $\mathbf{R}^3$) invariantes par le groupe des déplacements, les transformations galiléennes :

$$u(t,x) \mapsto u(x, x - at) + a,$$

où a est une constante, et par les changements d'échelle :

$$x \mapsto \lambda x, \quad t \mapsto \lambda^{1-h} t, \quad u \mapsto \lambda^h u,$$

où h est réel arbitraire et $\lambda > 0$. D'autre part, on sait, dans certains cas (en particulier avant la formation des chocs), montrer que les solutions des problèmes compressibles convergent, lorsque la compressibilité disparaît, vers les solutions des équations de Navier-Stokes. Enfin, on dispose de théorèmes d'existence et de méthodes de calcul différents et plus systématiques dans le cas des équations de Navier-Stokes que dans le cas des problèmes de fluides compressibles.

Les équations de Navier-Stokes mettent en jeu $n + 1$ inconnues dans un ouvert Ω de $\mathbf{R}^n$ (avec $n = 2$ ou 3), un champ de vecteurs $\vec{U} = (u_1, u_2, ..., u_n)$, et une pression p. Elles s'écrivent sous la forme :

$$(1) \quad \frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \nabla \vec{u} = f - \nabla p + \nu \Delta \vec{u}, \quad \nu \geqslant 0,$$

$$(2) \quad \nabla \cdot \vec{u} = 0$$

$$(3) \quad \vec{u}(x,t)|_{\partial\Omega} = 0 \quad \text{si } \nu > 0, \qquad \vec{u}(x,t) \cdot \vec{n}|_{\partial\Omega} = 0 \quad \text{si } \nu = 0$$

où $\vec{n}$ désigne la normale extérieure à la frontière de l'ouvert :

$$(4) \quad \vec{u}(x,0) = \vec{u}_0(x) \text{ donné.}$$

Le cas $\nu = 0$ correspond aux *équations d'Euler*, qui sont donc antérieures aux équations de Navier-Stokes, la contribution de Navier étant d'approcher la viscosité par le terme $\nu \Delta u$, $\nu > 0$. Ce terme, qui peut être très petit, est de l'ordre de $1/\mathrm{R}$ (où R est le nombre de Reynolds) : il permet cependant de simplifier considérablement l'analyse mathématique.

Comme dans le chapitre 1, $u(x, t)$ représente un champ de vitesse du fluide ; ainsi, en supposant que $u(x, t)$ est une fonction régulière, on est conduit à introduire les trajectoires des particules du fluide, données par les équations :

$$(5) \quad \dot{x}(t) = u(x(t), t), \quad x(0) = x_0.$$

Compte tenu de (3), le champ $u(x, t)$ est soit tangent au bord $\partial\Omega$ de Ω si $\nu = 0$, soit nul sur $\partial\Omega$; ainsi, les trajectoires $x(t)$ restent à l'intérieur de Ω et, pour t fixé, l'application de Ω dans Ω définie par l'équation (5) :

$$(6) \qquad x_0 \mapsto x(t)$$

est une bijection de Ω sur Ω. La relation (2) est équivalente, d'après un théorème dû à Liouville, au fait que l'application $x_0 \mapsto x(t)$ conserve les volumes.

L'équation de conservation de la masse s'écrit sous la forme :

$$(7) \quad 0 = \frac{\partial \rho}{\partial t} + \nabla(\rho u) = \frac{\partial \rho}{\partial t} + u \nabla \rho = \frac{d}{dt}\rho(x(t), t).$$

On en déduit que si la densité $\rho(x, t)$ est homogène et égale à 1 (pour fixer les idées) à l'instant zéro, elle reste constamment égale à 1. L'équation (1) n'est alors que l'équation classique de la mécanique, $\overrightarrow{f} = m\overrightarrow{\gamma}$, dans laquelle le premier membre représente l'accélération et le second membre l'ensemble des forces qui contribuent au mouvement ; f représente les forces extérieures comme la gravitation, $\nu \Delta u$ l'action des forces de viscosité et ∇p les forces de pression. Contrairement à ce qui se passe dans la théorie des gaz compressibles, où la pression est calculée à partir des autres inconnues (vitesse, température et densité), par une loi d'état, ici la pression n'est créée que par l'incompressibilité et on peut considérer ∇p comme un multiplicateur de Lagrange lié à la relation $\nabla u = 0$. On peut d'ailleurs éliminer ce terme, soit en utilisant le formalisme de l'analyse fonctionnelle, soit celui des opérateurs pseudodifférentiels. Dans le cas où $\Omega = \mathbf{R}^3$, on peut faire un calcul explicite. En prenant la divergence des deux membres de (1), on obtient :

$$(8) \quad \sum_{i,j} \frac{\partial}{\partial x_i}\left(u_j \frac{\partial u_i}{\partial x_j}\right) - \nabla f = \sum_{i,j} \frac{\partial u_i}{\partial x_j}\frac{\partial u_i}{\partial x_j} - \nabla f = -\Delta p.$$

La dernière équation donne p par convolution avec le noyau $\frac{1}{4\pi}\frac{1}{|x|}$. On obtient donc finalement :

$$(9) \quad -\nabla p = -\frac{1}{4\pi}\int \frac{\overrightarrow{x-y}}{|x-y|^3}\left(\sum \frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i}(y) - \nabla f(y)\right)dy.$$

Bien entendu, comme on l'avait indiqué, le second membre de (9) fait intervenir une intégrale singulière qui est un opérateur pseudodifférentiel d'ordre -1. Enfin, la condition aux limites exprime dans le cas $\nu > 0$ que la viscosité empêche tout mouvement sur les parois et, dans le cas $\nu = 0$, que le fluide est tangent aux parois.

On remarque ensuite que la condition $\nabla \cdot u = 0$ permet de donner une forme faible au terme d'advection. Plus précisément, pour toute fonction $\varphi \in (\mathcal{D}(\Omega))^n$, à divergence nulle, on a :

$$(10) \qquad (\nabla p, \phi) = 0$$

et :

$$\left(\sum u_i \frac{\partial u}{\partial x_j}, \varphi\right) = -\sum u_i u_j \frac{\partial \varphi_j}{\partial x_i} = -(u, u\nabla\varphi)$$

ainsi, on dira qu'une fonction est solution faible si elle est à divergence nulle et vérifie au sens des distributions l'équation :

$$(11) \quad \frac{d}{dt}(u, \varphi) - (u, u\nabla\varphi) - \nu(u, \Delta\varphi) = (f, \varphi),$$

pour toute fonction $\varphi \in (\mathcal{D}(\Omega))^n$ à divergence nulle.

Pour construire une solution faible, il est essentiel d'obtenir des suites u_ε qui appartiennent à $L^2(0, T \times \Omega)$, pour T fixé arbitrairement grand, et qui convergent presque partout pour pouvoir passer à la limite dans le terme $u^i_\varepsilon u^j_\varepsilon$. Multipliant l'équation (1) par u et intégrant par parties, on obtient (en supposant que u est une fonction régulière) la relation :

$$(12)\quad \frac{d}{dt}\int_\Omega \frac{1}{2}|u(x,t)|^2 dx + \nu\int_\Omega |\nabla u(x,t)|^2 dx = \int_\Omega (f, u)\, dx,$$

cette relation (12) exprime le bilan d'énergie et, pour ν positif, elle assure que l'expression :

$$\int_0^T \int_\Omega |\nabla u(x,t)|^2 dx\, dt$$

est uniformément bornée. On introduit donc les espaces de type Sobolev suivants :

$$\begin{aligned} H &= \{u \in (L^2(\Omega))^3 \,|\, \nabla . u = 0, u . n|_{\partial\Omega} = 0\},\\ V_0 &= \{u \in H^1_0(\Omega) \,|\, \nabla . u = 0\},\\ V_1 &= \{u \in H^1(\Omega) \,|\, \nabla . u = 0, u . n|_{\partial\Omega} = 0\}. \end{aligned}$$

Compte tenu de la relation $\nabla u = 0$, on peut définir, comme une distribution, la valeur de $u . n|_{\partial\Omega}$; on a alors les inclusions $V_0 \subset V_1 \subset H$; l'espace V_0 est fermé dans V_1 et V_1 est dense dans H. Appliquant l'inégalité (12), on peut prouver (ce résultat est, pour l'essentiel, dû à Leray) le théorème suivant :

Théorème 1. Pour tout $u_0 \in H$, le problème (1)-(4) admet une solution faible au sens suivant : pour tout $T > 0$, u appartient à $L^\infty(0, T ; H) \cap L^2(0, T ; V_0)$ et vérifie, pour toute fonction φ de V_0, la relation :

$$(13)\quad \frac{d}{dt}(u\,\varphi) - (u, u . \nabla\varphi) + \nu(\nabla u, \nabla\varphi) = (f, \varphi).$$

Le théorème 1 est valable pour tout T et en toute dimension d'espace (en particulier pour $n = 2$, 3, qui sont les cas physiques). Il caractérise la solution des équations de Navier-Stokes par une relation de type variationnel (comme pour les problèmes paraboliques) ; ainsi, les équations de Navier-Stokes se prêtent bien à des calculs numériques par éléments finis. Par contre, il y a une différence fondamentale entre le cas $n = 2$ et le cas $n = 3$. En dimension 2, on peut prouver que si f est régulière, alors u est aussi une fonction régulière, ce qui assure l'unicité de la solution. Par contre, en dimension 3, on ne sait pas prouver la régularité de la solution pour tout temps, et donc l'unicité, sauf si la donnée initiale $\|u_0\|_{V_0}$ est petite par rapport à la viscosité et si f est égale à 0. Dans ce cas, on démontre, par des méthodes de perturbation, l'existence d'une solution régulière pour tout temps.

Dans le cas $\nu = 0$, on ne dispose plus de l'estimation sur le gradient de la solution et ainsi on ne peut, en dimension 3, envisager l'étude d'une solution faible définie pour tout temps. Par contre, pour t petit, on peut appliquer à l'équation d'Euler un théorème de type Cauchy-Kovalevskaïa (ce résultat sous sa forme primitive est dû à Lichtenstein en 1930 et il a été amélioré par de nombreux auteurs, sans que sa nature soit vraiment modifiée). La limitation sur le temps provient du fait suivant : on établit des estimations a priori sur la solution (ou sur une solution approchée) en majorant une norme convenable par la solution de l'équation différentielle ordinaire $y' = Cy^2$ qui explose en un temps fini. Par contre, en dimension 2, on utilise le fait que le rotationnel (qui décrit le tourbillon) s'identifie à un vecteur perpendiculaire au plan du mouvement, ce qui, dans les équations, se traduit par la relation :

$$(14)\quad \nabla \wedge (u . \nabla u) = u . \nabla(\nabla \wedge u) = u . \nabla\,\omega,$$

où ω désigne le rotationnel de u. La relation (14) implique qu'en dimension 2, et en l'absence de force extérieure (cette hypothèse n'est introduite que pour simplifier) on a la relation :

$$(15) \qquad \frac{\partial \omega}{\partial t} + u \nabla \omega = 0,$$

qui exprime la conservation du tourbillon au cours du mouvement. La relation (15) permet de contrôler la norme de $u(x, t)$ dans l'espace V_1 dès que u_0 appartient à V_1. On peut ainsi prouver l'existence pour tout temps d'une solution faible du problème (1)-(4) en dimension 2 avec $\nu = 0$. De plus, si le tourbillon est très régulier à l'instant $t = 0$, on peut montrer qu'il va rester régulier et établir ainsi la régularité et l'unicité de la solution (ce résultat non trivial utilise la dispersion de paire, c'est-à-dire l'analyse de l'évolution de la distance entre deux particules de fluide ; il est dû à Wolibner vers 1930).

En parallèle avec les équations d'évolution, il est naturel de considérer des équations de Navier-Stokes stationnaires :

$$(16) \quad -\nu\Delta u + u\nabla u = f - \nabla p\,; \quad \nabla u = 0,$$

avec les conditions aux limites $u|_{\partial\Omega} = 0$ (avec $\nu > 0$). Il est alors facile de voir que le problème (16) admet toujours des solutions, et n'en admet qu'une seule si ν est grand.

Il reste autour des équations de Navier-Stokes un certain nombre de problèmes que nous allons commenter :

(I) Déterminer si, en dimension 3, les équations d'Euler présentent, comme les équations des fluides compressibles, des singularités au bout d'un temps fini. En effet, on ne dispose que d'une majoration d'une norme convenable de la solution et le fait que cette majorante devienne infinie ne donne aucune indication sur le comportement de la solution elle-même. Une raison fondamentale pour laquelle l'étude de l'apparition des singularités pour l'équation d'Euler est plus difficile que pour les équations des fluides compressibles est que les singularités ne résulteront pas du choc des molécules (à cause de la condition $\nabla u = 0$) mais de phénomènes beaucoup plus complexes qui, a priori, n'ont aucune raison d'être localisés sur des surfaces.

(II) Déterminer si, en dimension 3, les solutions des équations de Navier-Stokes (avec $\nu > 0$) présentent *vraiment* des singularités. Leray avait montré que si ces singularités existaient, elles résidaient sur un ensemble assez petit. Ce résultat a été progressivement amélioré et finalement Cafarelli, Kohn et Nirenberg ont montré que le support singulier de la solution de (1)-(4) est, pour tout $\nu > 0$, contenu dans un ensemble de mesure de Hausdorff au plus égale à 1 dans l'espace à quatre dimensions $\mathbf{R}_x^3 \times \mathbf{R}_t$.

(III) Déterminer si, en dimensions 2 et 3, les solutions des équations de Navier-Stokes avec $\nu > 0$ convergent vers les solutions des équations d'Euler lorsque la viscosité tend vers 0. Ce problème est résolu en dimension 2 si $\Omega = \mathbf{R}^2$ et en dimension 3 si $\Omega = \mathbf{R}^3$, si on se limite à un temps assez petit (en relation avec l'existence locale des solutions). Par contre, si le domaine présente une frontière, la condition aux limites $u|_{\partial\Omega} = 0$ crée, pour ν petit, une couche limite (région de fort tourbillon). Cette couche limite a une importance fondamentale dans les problèmes d'aérodynamisme (fig. 1). La non-linéarité du problème va entraîner une propagation de ces couches vers l'intérieur et donc, sur le plan mathématique, empêcher de contrôler la solution et de prouver sa

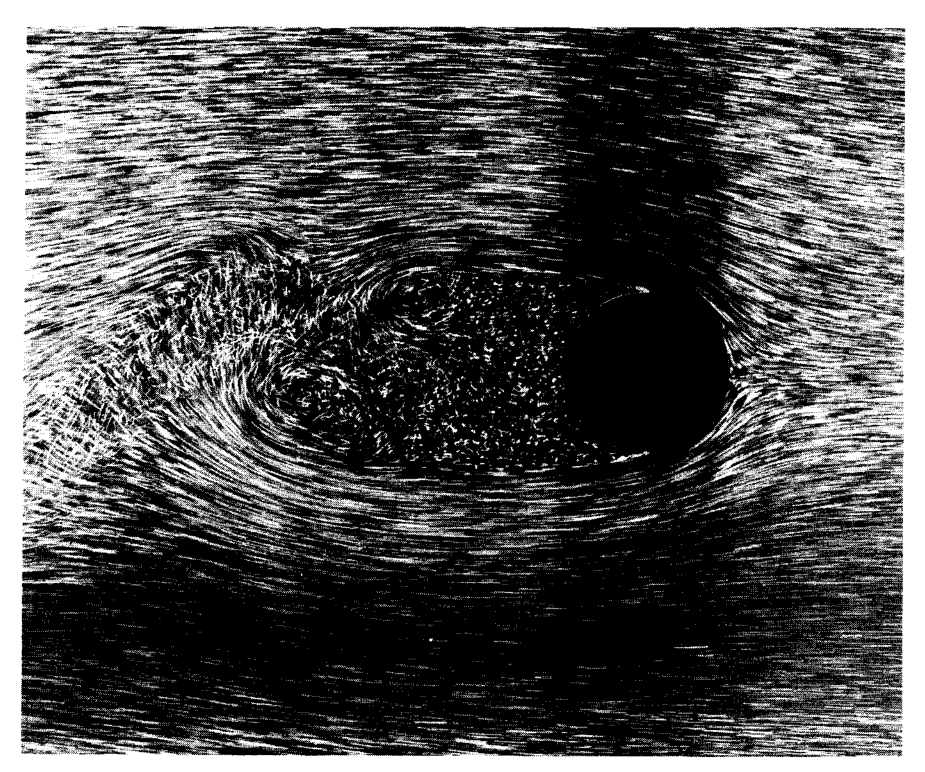


fig. 1 - *Cette photo met en évidence les lignes de courant d'un fluide autour d'un obstacle (nombre de Reynolds 2000). On remarque à la fois l'apparition de la couche limite et la propagation des tourbillons loin derrière l'obstacle (O.N.É.R.A.).*

convergence vers une solution de l'équation d'Euler.

(IV) Étudier la structure asymptotique, lorsque le temps augmente indéfiniment, des solutions de l'équation :

$$(17) \quad \frac{\partial u}{\partial t} - \nu \Delta u + u \, . \nabla u = f - \nabla p, \quad \nabla u = 0$$

en relation avec les solutions de l'équation stationnaire :

$$(18) \quad -\nu \Delta u + u \, . \nabla u = f - \nabla p, \quad \nabla u = 0.$$

Pour étudier les problèmes (17) et (18) dans un domaine borné Ω, avec la condition aux limites $u|_{\partial\Omega} = 0$, on utilise les propriétés de l'opérateur $-\nu\Delta$, en particulier le fait que sa résolvante est aussi compacte ; on peut ainsi montrer (Temam) que l'ensemble des solutions de (18) appartient en général à une variété analytique de dimension finie et que les solutions de (17) résident aussi arbitrairement près d'une variété analytique de dimension finie. On peut en conclure que l'étude asymptotique des solutions de Navier-Stokes peut se déduire de l'étude des solutions des systèmes différentiels ordinaires ; en particulier, on peut donner des conditions suffisantes sur ν et f pour que la variation de ν engendre une bifurcation de Hopf.

Les problèmes mathématiques concernant les équations de Navier-Stokes sont reliés à différentes interprétations de la notion de turbulence. Lorsqu'on augmente expérimentalement le nombre de Reynolds d'un fluide (par exemple en augmentant les

forces extérieures ou la vitesse des parois du récipient du fluide, ce qui par un changement d'inconnues se ramène à un problème de même nature), on voit le fluide passer d'un état régulier à un état turbulent.

L'état turbulent se caractérise par les propriétés suivantes.

Il apparaît des structures très complexes mais autosimilaires dans lesquelles le tourbillon prend des valeurs arbitrairement grandes. À petite échelle, ces structures sont homogènes, mais plus la taille du tourbillon augmente, plus les régions où celui-ci est grand sont petites. On dit que l'on a un phénomène d'*intermittence* (fig. 1).

Compte tenu de leur universalité, les équations de Navier-Stokes devraient contenir dans leurs solutions tous les phénomènes de turbulence, en liaison avec les pathologies (I) à (IV). Voici des explications possibles de la turbulence :

a) l'apparition de singularités au bout d'un temps fini pour les équations d'Euler en dimension 3 ;

b) comportement « turbulent » lorsque la viscosité tend vers 0 ;

c) une cascade de bifurcations dans les solutions de (18) qualitativement semblables aux bifurcations en dimension finie.

Il est conjecturé que ces trois phénomènes doivent avoir en commun un certain nombre de propriétés fondamentales, dont le fait de se produire sur des ensembles de dimension fractionnaire qui ressemblent à des ensembles de Cantor. En particulier, le problème (18) devrait exhiber, pour une valeur de ν assez grande, un « attracteur étrange ». Signalons, d'une part, que des bornes supérieures pour la mesure de Hausdorff de ces éventuels attracteurs ont été obtenues par plusieurs auteurs (Douady et Osterlé, en 1980) et que, d'autre part, il est possible de construire des modèles simples qui présentent un comportement asymptotique turbulent. Un des meilleurs candidats pour cette démarche est le système de Lorenz. Il est obtenu à partir du problème de Bénard : on considère un fluide incompressible dans un récipient bidimensionnel chauffé par en dessous ; les équations (17) et (18) sont alors couplées avec une équation pour la température et f se réduit aux forces de gravité. Lorenz développe alors u et f en série de Fourier et, ne conservant que trois des premiers coefficients, obtient le système :

$$(19) \qquad \begin{cases} \dot{x} = -x + y \\ \dot{y} = rx - y - xz. \\ \dot{z} = xy - bz \end{cases}$$

Bien entendu, il n'y a aucune justification mathématique dans cette démarche, mais il est frappant de constater les similitudes entre les comportements asymptotiques des solutions de (19) et ceux, observés expérimentalement, du problème de Bénard (fig. 2).

3. L'équation de Korteweg et de Vries

En 1865, Scott Russell observa sur un canal rectiligne une onde de surface créée par le choc de deux péniches, qu'il appela onde solitaire ; il fut frappé par la stabilité du phénomène et raconte qu'il put la suivre à cheval, à vitesse constante, pendant plusieurs kilomètres.

Pour expliquer ce phénomène, dit de *soliton*, on peut utiliser un système de deux équations à une dimension d'espace :

$$(1) \qquad \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(uh) = 0,$$

$$(2) \qquad \frac{\partial u}{\partial t} + uu_x + gh_x = 0,$$

dans lesquelles h désigne la hauteur de l'eau et u la vitesse du fluide, dans un canal



fig. 2 – *Expérience de Bénard : fumée de cigarette entre deux cylindres en rotation (O.N.É.R.A.).*

supposé indéfiniment long et peu profond ; g est une constante qui représente la gravité. Le système (1), (2) est un système hyperbolique non linéaire, du type évoqué dans le chapitre 2, et il s'obtient en écrivant les équations classiques des fluides incompressibles et en introduisant un paramètre petit lié à la profondeur du canal.

En effectuant un nouveau développement asymptotique, lié à l'amplitude de l'onde, Korteweg et de Vries obtinrent en 1895 l'équation :

$$\frac{\partial h}{\partial t} + \alpha h \frac{\partial h}{\partial x} + \beta \frac{\partial^3 h}{\partial x^3} = 0,$$

où α et β désignent des constantes. Par un changement de variables et d'inconnues, on peut écrire l'équation sous la forme :

$$(3) \qquad \frac{\partial u}{\partial t} - 6u \frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0.$$

Korteweg et de Vries observèrent que cette équation admet des solutions « ondes solitaires » de la forme :

$$(4) \quad u(x, t) = -\frac{c}{2}\left(ch^2 \frac{\sqrt{c}}{2}(x - ct)\right)^{-1}, \ c > 0 ;$$

ce sont des ondes qui se propagent avec la vitesse c sans se déformer. Cette situation est fondamentalement différente de celle

qui peut se produire dans un modèle linéaire. Les équations linéaires dont les solutions se propagent sont les équations hyperboliques ; les vitesses de propagation sont imposées par l'équation, et la forme de la solution qui se propage est indépendante de la vitesse. Dans cet exemple non linéaire, on peut avoir des ondes qui se propagent avec n'importe quelle vitesse positive, mais leur forme est imposée par la vitesse selon la formule (4). En particulier, plus la vitesse est grande, plus la solution est grande.

Intuitivement, on peut expliquer cette conservation de la forme par la compétition de deux phénomènes, en disant que l'équation (3) est un « mélange » des équations :

$$\frac{\partial u}{\partial t} + \frac{\partial^3 u}{\partial x^3} = 0, \tag{5}$$

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} = 0 \tag{6}$$

En multipliant (5) par u et en intégrant de $-\infty$ à $+\infty$, on observe que l'expression :

$$\int_{-\infty}^{+\infty} u^2(x, t)\, dx$$

reste constante, autrement dit la solution ne tend pas vers 0 dans l'espace $L^2(\mathbf{R})$; mais on remarque, en utilisant par exemple la transformation de Fourier, que, pour x fixé (et pour une donnée initiale assez régulière), $u(x, t)$ tend vers 0 comme $t^{-3/2}$ pour t tendant vers l'infini. L'équation (5) décrit une onde qui se disperse, tandis que l'équation (6) est l'équation de Burger, étudiée au chapitre 1, qui engendre un choc.

En fait, l'équation de Korteweg-de Vries possède des solitons car les phénomènes de création de chocs et de dispersion se compensent pour arriver à un état d'équilibre. En 1965, Kruskal, Miura et Zabusky remarquèrent une analogie entre le système (1) et le modèle de Fermi-Pasta-Ulam. Ce modèle avait été introduit en 1950 pour décrire la répartition de l'énergie sur un cristal non conducteur ; il était composé de 32 équations différentielles ordinaires non linéaires couplées et a été calculé par approximation numérique sur l'ordinateur Maniac I de Los Angeles. Les auteurs observèrent, au lieu d'une répartition uniforme de l'énergie sur tout le cristal, des phénomènes de périodicité ou de presque périodicité. Ces phénomènes furent expliqués beaucoup plus tard par Arnold, Moser et Kolmogorov dans le cadre de la théorie des systèmes hamiltoniens. Kruskal et d'autres entreprirent alors de discrétiser l'équation de Korteweg - de Vries et découvrirent deux types de phénomènes.

a) Dans le cas où on considère des données initiales $\varphi(x)$ périodiques en espace, la solution reste périodique en espace mais est une fonction presque périodique en temps.

b) Dans le cas où la donnée initiale $\varphi(x)$ est une fonction tendant vers 0 pour $|x| \to +\infty$ et si elle est « composée » d'une somme finie d'ondes solitaires, ordonnée suivant la taille (les plus petites d'abord), alors les plus grandes ondes rattrapent les plus petites et, au bout d'un certain laps de temps, on obtient à nouveau une solution soliton, avec la différence que, maintenant, ce sont les plus grandes qui sont en avant : l'ordre s'est inversé.

La démonstration de ce type d'observation a pu être entreprise d'après une idée de Lax (1968). Il introduit un opérateur auxiliaire défini par la relation :

$$H(t)\,\psi = -D^2\psi + u(x, t)\psi.$$

où D désigne ici la dérivation par rapport à x

L'opérateur H(t) dépend du paramètre t par l'intermédiaire de la fonction u. C'est l'opérateur autoadjoint (dans l'espace de Hilbert $L^2(\mathbf{R})$) usuel de la mécanique quantique. On cherche alors des conditions suffisantes pour que H(t) reste unitairement équivalent à lui-même, ce qui signifie qu'il existe une famille U(t) d'opérateurs linéaires unitaires de $L^2(\mathbf{R})$ sur lui-même tels que l'on ait :

$$H(t) = U(t)H(0)U^{-1}(t) = U(t)H(0)U^*(t).$$

Cela se produit, par exemple, si $u(x, t)$ est déformé par translation :

$$(7) \qquad u(x,t) = \varphi(x - t).$$

On remarque alors que $u(x, t)$ est solution de l'équation :

$$(8) \qquad \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0.$$

Plus généralement, soit L(t) une famille d'opérateurs antiadjoints dans $L^2(\mathbf{R})$, dépendant assez régulièrement du paramètre t (en particulier, leur domaine en tant qu'opérateur antiadjoint est indépendant de t), alors l'équation :

$$(9) \qquad \frac{\partial \varphi}{\partial t} = L(t)\,\varphi, \quad \varphi(0) = \varphi_0$$

admet une unique solution donnée par $\varphi(t) = U(t)\varphi_0$, où U est un opérateur unitaire. Dans le cas des équations (7) et (8), L(t) est l'opérateur $-d/dx$ et U(t) est le groupe des translations. Avec ce formalisme, on remarque par un calcul très simple que la relation :

$$U^*(t)H(t)U(t) = H(0)$$

est équivalente à la relation

$$(10) \quad 0 = \frac{d}{dt}(U^*(t)H(t)U(t)) = U^*(t)(-L(t)H(t) + H'(t) + H(t)L(t))U(t).$$

L'équation (10) s'écrit aussi sous la forme particulièrement commode :

$$(11) \quad \frac{dH}{dt} = L(t)H(t) - H(t)L(t) = [L, H].$$

Comme H(t) est l'opérateur $(d^2/dx^2) + u(x, t)$, $(dH)/(dt)$ est l'opérateur de multiplication par $(\partial u)/(\partial t)$.

Comme exemple d'opérateur moins trivial que $\partial/(\partial x)$, on peut introduire l'opérateur :

$$L(t) = \frac{\partial^3}{\partial x^3} + b\frac{\partial\,.}{\partial x} + \frac{\partial}{\partial x}(b\,.)\,;$$

L(t) est un opérateur antiadjoint dans $L^2(\mathbf{R})$; pour que le second membre de (11) soit un opérateur de multiplication scalaire, il faut et il suffit que b figurant dans la définition de L(t) et u figurant dans la définition de H(t) soient reliés par la relation $b = -(3/4)\,u$. On a alors :

$$[L, H]\varphi = \left(6\,u\frac{\partial u}{\partial x} - \frac{\partial^3 u}{\partial x^3}\right)\varphi,$$

et la relation (11) est alors équivalente à l'équation de Korteweg - de Vries. On a donc démontré le théorème suivant, dû à Lax : Pour que l'opérateur :

$$-\frac{\partial^2}{\partial x^2} + u(x,t)$$

soit unitairement équivalent à l'opérateur :

$$-\frac{\partial^2}{\partial x^2} + u(x,0)$$

par l'intermédiaire des opérateurs unitaires *définis* par

$$L(t) = \frac{\partial^3}{\partial x^3} + \frac{3}{4}\left(u\frac{\partial}{\partial x} + \frac{\partial}{\partial x}u\right),$$

il faut et il suffit que la fonction u soit solution de l'équation de Korteweg - de Vries. Le progrès accompli est dû en particulier au fait que, pour $|x| \to \infty$, L(t) est unitairement équivalent à $\partial^3/(\partial x^3)$.

D'autre part, on peut analyser les fonctions φ solutions de l'équation :

$$(12) \qquad -\frac{d^2\varphi}{dx^2} + u\,\varphi = \lambda\,\varphi$$

ou, ce qui revient au même, faire l'analyse spectrale de l'opérateur :

$$-\frac{d^2\varphi}{dx^2} + u\,\varphi.$$

On trouve deux types d'objets. Pour $\lambda < 0$, il existe une suite de nombres $\lambda_1 < \lambda_2 < \dots < \lambda_n < 0$ pour lesquels (12) admet une solution φ non identiquement nulle et de carré sommable. Ces nombres correspondent aux valeurs propres ou aux états liés du système. Bien entendu, si $H(t)$ reste unitairement équivalent à $H(0)$, ces nombres sont constants. Les vecteurs propres correspondants φ_k se comportent asymptotiquement, pour $|x| \to \infty$, comme :

$$C_k \pm \exp(-\lambda_k |x|).$$

On choisit C^+ tel que φ_k soit un vecteur de norme 1 dans $L^2(\mathbf{R})$. Cette construction a engendré une famille de couples $(\lambda_k, C_k^+(t))$. Ensuite, on remarque que, pour $\lambda = \xi^2$, on peut, pour tout t, construire une solution oscillante φ_ξ de (12) dont le comportement asymptotique est donné par :

$$\Phi_\xi = \begin{cases} e^{-ix\xi} + R(\xi)\,e^{ix\xi} \\ T(\xi)e^{-ix\xi}. \end{cases}$$

Les lettres R et T réfèrent aux mots réflexion et transmission en liaison avec l'interprétation physique de cette construction.

On procède donc de la manière suivante : à tout opérateur de la forme $-D^2 + u(x, t)$, on associe les éléments $(\lambda_k, C_k^+(t)), (T(\xi), R(\xi, t)),\ \ \xi \in \mathbf{R}$. Cet ensemble s'appelle les données de scattering. On utilise alors un théorème de Gelfand-Levitan-Marchenko qui permet de déterminer le potentiel $u(x, t)$ à partir des données de scattering par une équation intégrale assez explicite. Le progrès effectué est le suivant. Comme, pour $|x| \to \infty$, $L(t)$ se réduit à D^3, le calcul du comportement asymptotique des solutions de (12) peut être fait explicitement en fonction des données de scattering à l'instant zéro ; on détermine les données de scattering à l'instant t par les relations :

$$(13) \qquad \begin{cases} \lambda_k(t) = \lambda_k(0), \\ C_k(t) = C_k(0)\exp[4(-\lambda_k)^{3/2}t], \\ T(\xi, t) = T(\xi, 0), \\ R(\xi, t) = R(\xi, 0)\exp(8i\xi^3 t). \end{cases}$$

On obtient ensuite le potentiel $u(x, t)$ à l'instant t. L'ensemble de ce programme s'appelle la méthode inverse, et il fournit les outils qui permettent de démontrer les résultats observés numériquement par Kruskal, Miura et Zabusky dans le cas de données initiales tendant vers zéro à l'infini.

Dans le cas de problèmes périodiques, la situation est beaucoup plus compliquée, mais, en 1976, McKean et Trublowitz ont démontré que toutes les solutions $u(x, t)$ de l'équation de Korteweg - de Vries périodiques en espace étaient des fonctions quasi périodiques par rapport au temps, ce qui confirme les expériences numériques de Kruskal et d'autres. Tout autant que les résultats, l'esprit de ces démonstrations et les analogies qu'ils suggèrent sont fondamentales.

On remarque que les solutions du système peuvent s'écrire, en prenant les logarithmes des solutions des équations, sous la forme :

$$(\lambda_k(0), \ln C_k(0) + \beta_k t),$$
$$(T(\xi, 0), \ln|R(\xi, 0)| + \gamma(\xi)t).$$

Ce sont des fonctions linéaires de t ; on dit que l'on a globalement intégré l'équa-

tion de Korteweg - de Vries. Cette situation est formellement identique au *théorème de Liouville,* dont nous rappelons l'énoncé. On considère dans $\mathbf{R}^{2n}$, ensemble des couples (p, q), un système hamiltonien :

$$(14) \qquad \frac{\partial p}{\partial t} = -\frac{\partial H}{\partial q}, \quad \frac{\partial q}{\partial t} = \frac{\partial H}{\partial p},$$

et on dit qu'une famile $E_k(p, q)$ de fonctions, $1 \leqslant k \leqslant m$, forme un système linéairement indépendant en involution si les deux conditions suivantes sont satisfaites :

– l'application tangente :

$$(p, q) \mapsto (E_1, E_2, ..., E_m)$$

de $\mathbf{R}^{2n}$ dans $\mathbf{R}^m$ est de rang m ;

– pour tout couple k, l d'indices, on a la relation suivante :

$$(15) \qquad \sum_{i=1}^{n} \left(\frac{\partial E_k}{\partial p_i} \frac{\partial E_l}{\partial q_i} - \frac{\partial E_l}{\partial p_i} \frac{\partial E_k}{\partial p_i} \right) = 0.$$

On a alors le résultat suivant : tout système hamiltonien admettant n intégrales premières en involution et linéairement indépendantes est complètement intégrables.

Comme dans l'énoncé du théorème de Liouville, on dispose d'une infinité de fonctionnelles invariantes pour l'équation de Korteweg - de Vries, dont les premières :

$$i_0(u) = \int_{-\infty}^{+\infty} u(x, t)\, dx,$$

$$i_1(u) = \int_{-\infty}^{+\infty} \frac{1}{2} u^2(x, t)\, dx,$$

$$I_2(u) = \int_{-\infty}^{+\infty} (u^3(x, t) + \frac{1}{2} u_x^2(x, t))\, dx,$$

$$I_3(x) = \int_{-\infty}^{+\infty} (u^4(x, t) + 2uu_x^2 + \frac{1}{5}(u_{xx})^2)\, dx,$$

avaient été déterminées par Kruskal et ses collaborateurs avant l'introduction de la méthode inverse.

D'autre part, l'équation de Korteweg - de Vries est équivalente à l'équation :

$$(16) \qquad \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(\frac{\delta I_2}{\delta u} \right)$$

où $\delta I_2/\delta u$ désigne la dérivée au sens de Gateaux de la fonctionnelle $I_2(u)$, dérivée qui s'obtient par des méthodes standards de calcul des variations. On peut ensuite munir l'espace $H^1(\mathbf{R})$ d'une structure algébrique qui généralise la notion d'espace symplectique. Pour cette structure, (16) est une équation hamiltonienne et les fonctionnelles I_n sont en involution. Ainsi, l'équation de Korteweg - de Vries fournit un exemple qui satisfait à la fois aux hypothèses et aux conclusions du théorème de Liouville en dimension infinie. Un tel théorème n'est pas encore démontré (la difficulté essentielle étant de caractériser le fait que l'ensemble des fonctionnelles invariantes est assez gros).

La détermination successive des invariants de l'équation de Korteweg - de Vries a conduit à lui associer une structure algébrique très complexe qui est reliée à la géométrie algébrique (algèbres de Kac-Moodie), néanmoins, il est facile de montrer comment faire intervenir des variétés algébriques et des fonctions analytiques dans la théorie de Korteweg - de Vries. On remarque pour cela que la fonction :

$$u(x, t) = \frac{-c/2}{\mathrm{ch}^2 \frac{1}{2} c(x - ct)},$$

définie pour x réel, se prolonge en une fonction méromorphe dans le plan complexe. Cette remarque, jointe à la perspective d'utiliser le théorème de Liouville en dimension finie, conduit à chercher des

solutions rationnelles de l'équation de Korteweg - de Vries de la forme :

$$u(x,t) = 2\sum_{j=1}^{N}\frac{r_j(t)}{(x-x_j(t))^2}.$$

En reportant le second membre de cette équation dans l'équation de Korteweg - de Vries écrite sous la forme :

$$(17) \qquad u_t - 3uu_x + \frac{1}{2}u_{xxx} = 0,$$

on trouve que, pour que u soit solution de (17), il faut et il suffit que $r_j(t)$ soit identique à 1 et que les $x_j(t)$ vérifient le système différentiel ordinaire :

$$(18) \quad \dot{x}_j = 6\sum_{k\neq j}(x_j-x_k)^{-2}, \quad 1 \leqslant j \leqslant N,$$

avec la condition :

$$(19) \quad \sum_{k\neq j}(x_j-x_k)^{-3} = 0, \quad 1 \leqslant j \leqslant N.$$

En s'inspirant des résultats de Lax, Kruskal, Garner et Als, on montre alors que la solution de (18) coïncide avec la solution du système hamiltonien défini par :

$$(20) \qquad \frac{dx}{dt} = -\frac{\partial H}{\partial y}, \quad \frac{dy}{dt} = \frac{\partial H}{\partial x};$$

$$H(x,y) = \frac{1}{3}\sum y_j^3 + 9\sum_{j\neq i} y_i(x_i-x_j)^{-2}.$$

Il n'est pas évident que le système (18), (19) ait une solution ; la difficulté réside dans la relation (19). En effet, il faut vérifier que si cette relation est vérifiée pour $t = 0$, elle est aussi vérifiée lorsque $x(t)$ varie avec t. On montre que cela impose aux x_i d'appartenir au plan complexe et à N d'être de la forme $N = (n(n+1))/2$, pour n entier. Comme dans le cas continu, on met alors en évidence N intégrales premières en involution et on conclut que l'on obtient un système hamiltonien totalement intégrable sur la variété complexe définie par :

$$\sum_{k\neq j}(x_j-x_k)^{-3} = 0, \quad 1 \leqslant j \leqslant N.$$

Pour vérifier que :

$$2\sum_{j=1}^{N}\frac{1}{(x-x_j(t))^2}$$

est solution de l'équation de Korteweg - de Vries (18), on utilise essentiellement la formule d'addition :

$$\Phi(x)\,\Phi'(y) + \Phi(x)\,\Phi(y) = \Phi(x-y)$$
$$[\Phi'(x) + \Phi'(y)] - \Phi'(x-y)\,[\Phi(y) - \Phi(x)],$$

valable pour la fonction :

$$\Phi(x) = x^{-2},$$

mais aussi pour toute fonction elliptique.

C'est pourquoi on peut reprendre la théorie en cherchant des solutions de l'équation de Korteweg - de Vries sous la forme :

$$u(x,t) = \sum_{j=1}^{N}\mathfrak{p}\Big(\frac{1}{2}(x-x_j(t))\Big),$$

où $\mathfrak{p}$ est une fonction elliptique quelconque.

Les relations (18) et (19) donnent ici :

$$(21) \quad \dot{x}_j = 3\sum_{k\neq j}\mathfrak{p}\Big(\frac{1}{2}(x_j-x_k)\Big), \quad 1 \leqslant j \leqslant N,$$

$$(22) \quad \sum_{k\neq j}\mathfrak{p}'\Big(\frac{1}{2}(x_j-x_k)\Big) = 0, \quad 1 \leqslant j \leqslant N.$$

Le première difficulté est l'étude de l'invariance de la relation (22) sous l'action du système différentiel défini par (21). Il n'existe pas de résultats généraux et une telle variété invariante peut souvent être

vide. Dans le cas de la fonction elliptique définie par :

$$(\wp')^2 = 4\wp^3 - g_2\wp - g_3,$$

on sait, pour $N = 3$, analyser complètement la variété invariante définie par (21) et (22).

L'étude de cet exemple peut paraître très particulière en raison de l'intégrabilité complète de l'équation de Korteweg - de Vries et de tous les calculs exacts qui ont pu être menés à bien. Pourtant, cet exemple est fondamental car une partie, et parfois la totalité, des méthodes décrites ici s'appliquant à d'autres équations qui interviennent aussi bien en physique qu'en mathématiques « pures ». Sans les expliciter, on peut citer les suivantes : l'équation de Korteweg - de Vries modifiée intervient dans l'étude des ondes d'Alfvén et des plasmas froids, l'équation de Schrödinger non linéaire dans l'étude de la focalisation des faisceaux laser, l'équation de Sine-Gordon dans les ondes de spin et l'optique non linéaire, l'équation de Boussinesq en hydrodynamique et en théorie des plasmas. L'équation de Kandomstev et Petviaschvili (qui est l'exemple principal pour lequel la théorie s'étend à plus d'une variable d'espace) intervient dans la description des milieux faiblement dispersifs.

4. Les équations de réaction-diffusion

On a vu au chapitre 2 que l'étude du comportement asymptotique des solutions de l'équation de Navier-Stokes était encore très fragmentaire. En particulier, il n'est pas possible de démontrer pour les équations de Navier-Stokes des résultats qualitatifs aussi précis que ceux que l'on observe sur des modèles à un nombre fini de degrés de liberté comme le système de Lorenz. Par contre, pour les équations de type Korteweg - de Vries, la méthode inverse a fourni une description complète du comportement asymptotique. Ce chapitre est consacré aux équations de réaction-diffusion pour lesquelles le comportement asymptotique est encore le problème essentiel. Pour ce type d'équation, on ne dispose pas de méthode inverse. On n'aura, en général, qu'un nombre fini d'ondes solitaires ; mais le rapport de parenté avec les équations différentielles ordinaires est bien plus importante et on peut obtenir, dans certains cas, des démonstrations complètes ; on s'appuie en particulier sur la théorie des systèmes dynamiques.

Les équations (ou systèmes) de réaction-diffusion s'écrivent sous la forme :

$$(1) \qquad \frac{\partial u}{\partial t} = \mathrm{D}(u)\,\Delta u + \mathrm{F}(u),$$

où $u(x, t)$ est une fonction vectorielle à valeurs dans $\mathbf{R}^m$ définie pour la variable x parcourant un ouvert Ω de $\mathbf{R}^n$. Lorsque Ω est différent de $\mathbf{R}^n$, on suppose que $u(x, t)$ vérifie sur le bord des conditions aux limites classiques. Dans cette équation, F est une fonction non linéaire régulière définie dans $\mathbf{R}^n$ et à valeurs dans $\mathbf{R}^n$; Δ désigne le laplacien usuel et $\mathrm{D}(u)$ est une matrice symétrique positive ou définie positive. Lorsque D est définie positive, le problème est non linéaire et parabolique, lorsque D n'est pas définie positive, on est en présence d'un problème parabolique dégénéré. Dans tous les cas, des méthodes de perturbation permettent de prouver que, pour toute donnée $u_0(x)$ définie à l'instant $t = 0$, il existe au moins, pour t petit et positif, une solution du sytème (1) qui vérifie $u(x, 0) = u_0(x)$.

Ces équations interviennent dans la description de phénomènes non linéaires

dans lesquels la dépendance en espace introduit une évolution de type mouvement brownien. On les rencontre dans la modélisation des réactions chimiques, et, en particulier, dans les phénomènes de combustion, lorsque la vitesse de propagation de la flamme est assez lente par rapport à la cinétique chimique, contrairement au régime de détonation qui, lui, relève des systèmes hyperboliques décrits dans le chapitre 1. On les rencontre aussi dans l'analyse des facteurs intervenant dans la propagation de l'influx nerveux et dans la dynamique des populations.

Les théorèmes d'existence globale, et certains résultats asymptotiques, reposent sur la notion de région invariante, que l'on va décrire sur un exemple simple. Supposons, pour simplifier, que $\Omega = \mathbf{R}$; on considère des solutions $u(x, t)$ qui tendent vers des limites finies u^+ et u^- lorsque x tend vers plus ou moins l'infini. On dira qu'une région $\mathcal{Q}$ de $\mathbf{R}^m$ (contenant u^+ et u^-) est invariante si l'appartenance $u_0(x) \in \mathcal{Q}$ pour tout x réel entraîne, pour $t > 0$ (éventuellement petit), la relation $u(x, t) \in \mathcal{Q}$. La notion de région invariante conduit à étudier, pour t fixé, le comportement des courbes $x \longmapsto u(x, t)$ dans $\mathbf{R}^m$. C'est la généralisation aux équations aux dérivées partielles de l'analyse faite, pour les équations différentielles ordinaires, à l'aide du plan des phases. On a par exemple le théorème suivant :

Supposons que $D(u)$ est une matrice diagonale. Alors, tout cube :

$$\mathcal{Q} = \prod_{i=1}^{m}]a_i, b_i[, \quad a_i < b_i,$$

de $\mathbf{R}^m$ à faces parallèles aux hyperplans de coordonnées, tel que, sur chaque face, le champ F soit rentrant, est une région invariante. En effet, si la courbe $u(x, t)$ venait de sortir de $\mathcal{Q}$ il y aurait une valeur limite (x^*, t^*) pour laquelle $u(x, t)$ atteindrait la frontière de $\mathcal{Q}$, par exemple sur la face $x_i = b_i$. Cela entraînerait que $u_i(x^*, t^*)$ atteindrait son maximum au point (x^*, t^*). On écrit alors la i-ième composante de l'équation (1) pour obtenir :

$$(2) \qquad \frac{\partial u_i}{\partial t} - D_i(u)\frac{\partial^2 u_i}{\partial x^2} - F_i(u) = 0.$$

Comme x^*, t^* est un maximum de u_i pour $x \in \mathbf{R}$ et $0 < t < t^*$, on a :

$$\frac{\partial u_i}{\partial t}(x^*, y^*) \geqslant 0 \quad \text{et} \quad \frac{\partial^2 u_i}{\partial x^2}(x^*, t^*) \leqslant 0.$$

Comme F rentre dans $\mathcal{Q}$ et $u_i(x^*, t^*) = b_i$, on a $F_i(u) < 0$; ainsi, le premier terme de (2) est une somme de termes positifs dont un au moins strictement, ce qui conduit à une contradiction. De plus, une analyse plus précise permet de montrer que :

$$\mathcal{Q} = \prod_{i=1}^{m}]a_i, b[$$

est une région invariante même si F, au lieu d'être strictement rentrant, est rentrant ou tangent.

L'existence d'une région invariante compacte permet d'obtenir pour la solution de (1) une majoration uniforme dans $L^\infty(\Omega)$. On en déduit que, pour $t < T$, $(du/dt) - \Delta u$ est uniformément bornée dans $L^\infty(\Omega)$ et donc, d'après les propriétés des problèmes paraboliques, très régulière. Il en résulte ensuite que la solution est définie et régulière, non seulement pour t petit, mais pour tout $t > 0$.

On se propose dans ce qui suit de décrire trois exemples significatifs.

Le premier est l'équation scalaire :

$$(3) \quad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - u(\alpha - u)(1 - u), \quad 0 < \alpha < 1.$$

L'expression explicite :

$$f(u) = -u(\alpha - u)(1-u)$$

n'est donnée que pour simplifier les calculs. En fait, comme on peut facilement le constater, seules interviennent les propriétés géométriques globales de f. Il convient de comparer l'équation (3) à l'équation différentielle ordinaire :

$$\frac{du}{dt} = f(u) = -u(\alpha - u)(1-u).$$

On remarque que 0, α et 1 sont des points stationnaires pour l'équation (4), que 0 et 1 sont des points stables, tandis que α est un point stationnaire instable. De plus, l'intervalle [0, 1] est une région invariante pour l'équation aux dérivées partielles (3) (dans ce cas, la forme classique du principe du maximum est suffisante). Cette équation a été introduite par Fischer en 1937 et est utilisée pour décrire des problèmes de génétique des populations ou de propagation de la flamme (combustion d'une mèche), il est naturel de chercher à décrire l'existence de fronts et leur stabilité, ce qui conduit à chercher des solutions de la forme $u(x, t) = u(x - ct)$. Reportant dans l'équation (3), on obtient l'équation différentielle ordinaire :

$$(4) \qquad -cu' = u'' - u(\alpha - u)(1-u),$$

qui s'écrit comme un système en introduisant la variable $v = u'$:

$$(5) \quad \begin{cases} u' = v \\ v' = -cv + u(\alpha - u)(1-u) \\ \qquad = -cv - F(u). \end{cases}$$

Un front, ou onde solitaire, est alors une solution de (5) qui vérifie les relations :

$$(6) \quad \lim_{\xi \to +\infty} u(\xi) = 1, \quad \lim_{\xi \to -\infty} u(\xi) = 0, \quad \lim_{|\xi| \to +\infty} v(\xi) = 0.$$

On supposera désormais que la fonction $F(u)$ vérifie la relation :

$$(7) \qquad \int_0^1 F(u)\,du > 0,$$

et on définit le point M par la relation :

$$(8) \qquad \alpha < M < 1, \quad \int_0^M F(u)\,du = 0.$$

Le point (0, 0) est un point hyperbolique pour le système (5) et il y a donc une solution de (5) vérifiant $\lim_{\xi \to -\infty} (u(\xi), v(\xi)) = 0$, tangente à la direction hyperbolique répulsive. Lorsque ξ augmente, v augmente jusqu'à ce que $(u(\xi), v(\xi))$ rencontre la courbe $cv = -F(u)$. Puis v diminue ; on obtient alors deux possibilités. Soit $(u(\xi), v(\xi))$ recoupe l'intervalle $[\alpha, 1]$ en un point $M_c > M$, soit $(u(\xi), v(\xi))$ recoupe la courbe $cv = -F(u)$ en un point $u > 1$; ces deux circonstances dépendent du choix de c. La première se produit pour c petit et la seconde pour c grand. On montre alors que le point M_c est une fonction croissante de c et qu'il existe donc une unique valeur $c^* > 0$ telle que $M_{c^*} = 1$; elle correspond au front, orbite de (5), connectant les points hyperboliques (0, 0) et (1, 0) (fig. 3).

De plus, il est clair que si $u(x - c^* t)$ est une onde solitaire, il en est de même de toute solution translatée. On peut alors montrer, en utilisant des théorèmes de comparaison pour les équations paraboliques non linéaires, le résultat de stabilité suivant, dû à Fife et Mac Leod : Supposons que la donnée initiale $u_0(x)$ soit majorée et minorée par deux ondes solitaires translatées ; plus précisément, il existe A et B positifs, éventuellement grands, tels que l'on ait :

$$u(x - A) \leqslant u_0(x) \leqslant u(x + B)$$

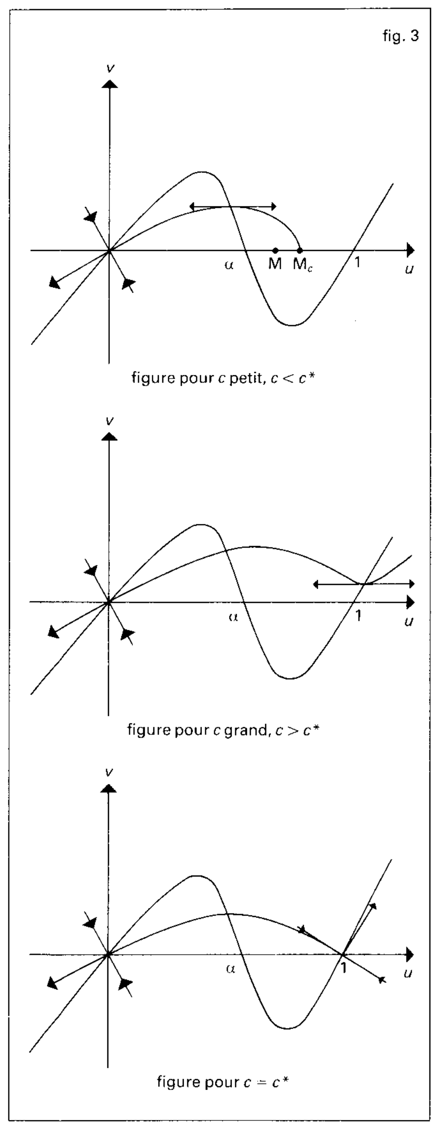


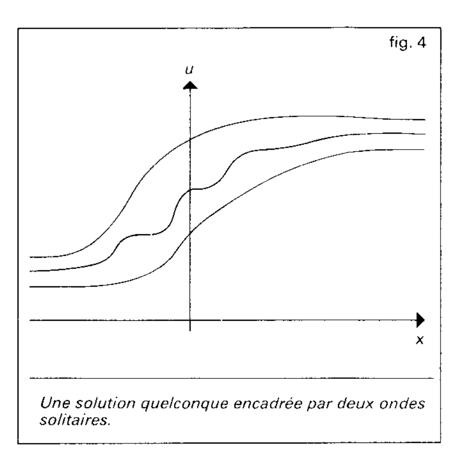
figure pour c petit, $c < c^*$

figure pour c grand, $c > c^*$

figure pour $c = c^*$

pour tout x réel. Alors, il existe un nombre D tel que $u(x, t)$, solution de (5), converge vers $u(x - c^*t - D)$ (fig. 4) .

La situation décrite dans ce type de problème est donc radicalement différente de celle qu'on rencontre pour les équations de Kortweg - de Vries. Il n'y a qu'une onde



Une solution quelconque encadrée par deux ondes solitaires.

solitaire (à une translation près) et sa vitesse est caractéristique du problème. On a pu démontrer son existence et sa stabilité en utilisant les techniques des équations différentielles ordinaires et la monotonie.

Le problème (3) est particulièrement simple, car il est scalaire ; cela correspond à des situations très simples dans lesquelles une seule quantité (ou le rapport de deux quantités) intervient. Dans le cas des systèmes, la situation est bien plus complexe.

Le deuxième exemple que nous nous proposons d'examiner est le système de Fitzugh et Nagumo. Dans leurs travaux sur la propagation de l'influx nerveux, Hodgkin et Huxley observèrent les phénomènes suivants :

a) il existe un seuil d'excitation : toute excitation inférieure à ce seuil ne produit pas de phénomène visible ;

b) au-dessus du seuil d'excitation, on obtient un signal qui a une forme constante et se propage à vitesse constante.

Hodgkin et Huxley ont conjecturé que ce comportement était dû à la diffusion et à l'interaction non linéaire entre un potentiel électrostatique et les réactions entre eux de plusieurs ions, et donc que ce

mécanisme était contrôlé par un système d'équations de réaction-diffusion. Le travail mathématique a consisté alors à montrer que les solutions du système d'Hodgkin-Huxley avaient un comportement conforme à l'expérience (existence de seuil et apparition d'ondes solitaires stables). Sans en démonter la validité, cela rend plausibles les équations de Hodgkin-Huxley. Ces propriétés ont été démontrées par C. Conley et plusieurs de ses élèves en utilisant une généralisation en dimension infinie des théorèmes de perturbation et de l'indice de Morse utilisés dans la théorie des équations différentielles ordinaires. Il n'est bien sûr pas possible de décrire ici ces travaux, mais on peut donner certaines idées en remplaçant le système de Hodgkin-Huxkley par un système plus simple, ayant les propriétés asymptotiques équivalentes, dû à Fitzugh et Nagumo :

$$(9) \qquad \frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} + f(v) - u,$$

$$(10) \qquad \frac{\partial u}{\partial t} = \varepsilon(\sigma v - \gamma u).$$

Comme dans l'exemple précédent $f(v)$ a le comportement qualitatif de la cubique :

$$f(v) = -v(v-a)(v-b), \quad 0 < a < b\,;$$

les nombres σ, γ, et ε sont des constantes positives et la droite $\sigma v - \gamma u = 0$ rencontre la courbe $u = f(v)$ uniquement au point 0.

On observe en premier lieu l'existence de deux familles de régions invariantes, celles qui sont limitées par des *grands* rectangles R et celles qui sont limitées par des *petits* rectangles $\tilde{R}$. Sur la figure 5 on a représenté un rectangle R et un rectangle $\tilde{R}$ en indiquant par des flèches la direction du champ $(f(v) - u, \sigma v - \gamma u)$. On désigne alors par R_+ le plus petit rectangle de la famille R et par R_- le plus grand rectangle de la famille $\tilde{R}$. Reprenant les théorèmes de comparaison qui ont conduit à la notion de région invariante, on démontre facilement les résultats suivants :

– pour toute donnée initiale $(v_0(x), u_0(x))$ définie sur $\mathbf{R}$, continue, et tendant vers 0 pour $|x| \to \infty$, le système (9), (10) admet une unique solution $(v(x, t), u(x, t))$ définie pour tout $t > 0$: pour t assez grand, la courbe $x \mapsto (v(x, t), u(x, t))$ est contenue dans le rectangle R_+ ;

– de plus, si la courbe $x \mapsto (v_0(x), u_0(x))$ est contenue dans R , la solution $(v(x, t), u(x, t))$ converge exponentiellement vers 0 lorsque t tend vers l'infini.

On a ainsi déjà mis en évidence deux notions : d'une part une région globalement attractante (dans le plan des phases) et d'autre part un bassin d'attraction de (0, 0) qui correspond au phénomène de *seuil* si l'excitation est trop petite : pour $(v_0(x), u_0(x)) \in R_-$ pour tout $x \in \mathbf{R}$, la solution tend exponentiellement vers 0.

On remarque ensuite que (0, 0) est un attracteur global pour l'équation :

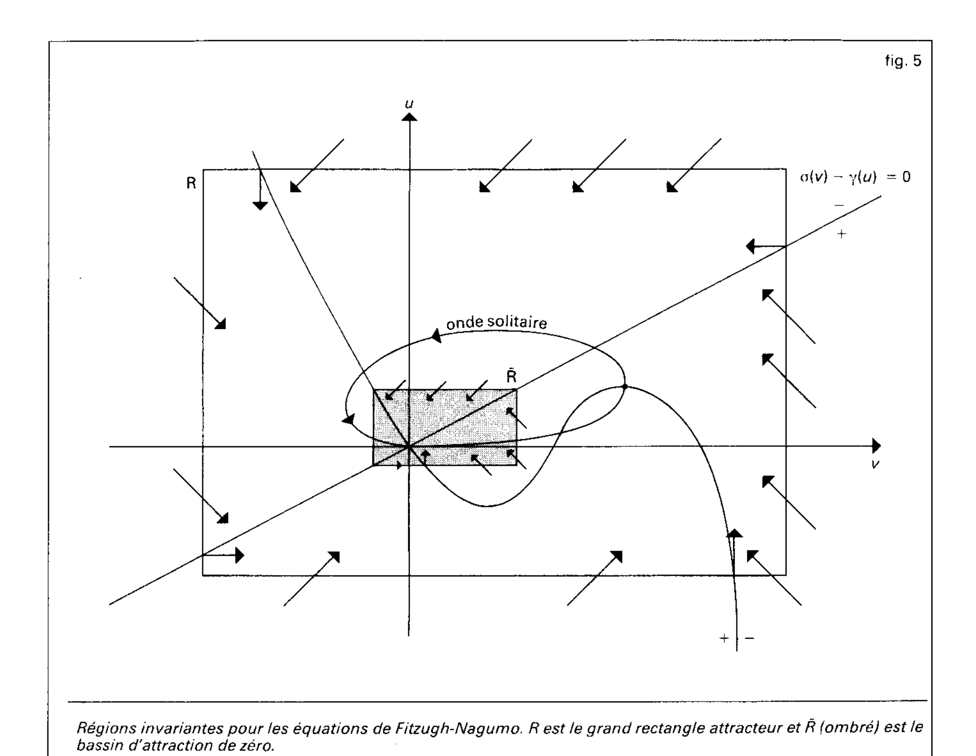
$$\frac{\partial v}{\partial t} = f(v) - u, \quad \frac{\partial u}{\partial t} = \varepsilon(\sigma v - \gamma u)\,;$$

mais l'apparition du terme de viscosité $\partial^2 v / \partial x^2$ modifie la première équation et, pour $v(x, t)$ (pour t fixé) introduit, là où cette courbe est convexe, un facteur de répulsion qui permet à la courbe de rester à l'extérieur du petit rectangle (fig. 5). On obtient une solution rigoureuse en cherchant des solutions de la forme :

$$(v(x, t), u(x, t)) = (v(x - ct), u(x - ct)),$$

ce qui, en introduisant la variable $w = dv/d\xi$, où $\xi = x - ct$, conduit au système différentiel ordinaire :

$$(15) \qquad v' = w, \quad -cu' = \varepsilon(\sigma v - \gamma u), \quad w' = -cw + u - F(u).$$



Régions invariantes pour les équations de Fitzugh-Nagumo. R est le grand rectangle attracteur et R̃ (ombré) est le bassin d'attraction de zéro.

On étudie alors, par des méthodes de perturbation, le système (15) pour ε petit. Le choix de cette situation (ε petit) se justifie en disant que les deux phénomènes, variation du potentiel v, et variation de u (réaction chimique) se produisent à des échelles de temps très grandes l'une par rapport à l'autre. On peut alors, pour ε petit, prouver l'existence d'une onde stationnaire stable.

Le troisième exemple que nous allons décrire très brièvement est la réaction de Bielouzov-Zabotinski. En 1959, Bielouzov a découvert, dans l'oxydation de l'acide malonique par le bromate de potassium, en présence d'ions cérium, des mécanismes de structuration spatiotemporels autoentretenus qui se présentent comme une superposition d'ondes solitaires radiales (fig. 6). L'analyse des mécanismes de réaction conduit à écrire un système à trois inconnues principales : $X \equiv HBrO_2$, $Y \equiv Br^-$ et Z forme oxydée de l'ion cérium.

On a alors le système :

$$(16)\quad \begin{aligned} \frac{\partial X}{\partial t} &= D\frac{\partial^2 X}{\partial x^2} + \alpha(Y - XY + X - \beta X^2) \\ \frac{\partial Y}{\partial t} &= \frac{1}{\alpha}(\gamma Z - Y - XY) \\ \frac{\partial Z}{\partial t} &= \delta(X - Z) \end{aligned}$$

où D, α, β, γ, δ désignent des constantes positives. Comme X, Y, Z désignent des concentrations de produits chimiques, il est naturel de décider qu'on se limite aux solutions du système (16) positives et bornées. En fait, on montre facilement que si a, b, c vérifient les inégalités :

$$a > \max\left(1, \frac{l}{\beta}\right), \quad c > a, \quad b > \gamma c,$$

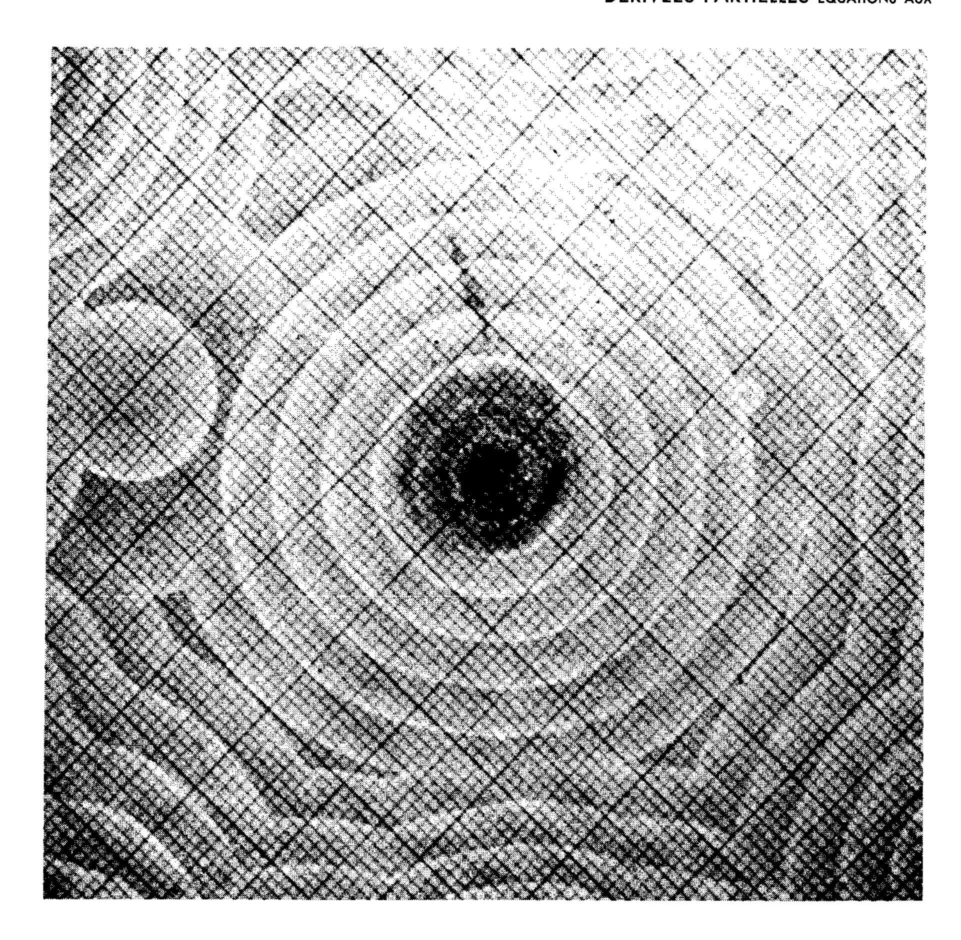


fig. 6 – *Réaction de Bielousov-Zabotinski (Society for Industrial and Applied Mathematics).*

la région :

$$R = \{0 \leqslant X \leqslant a, \quad 0 \leqslant Y \leqslant b, \quad 0 \leqslant Z \leqslant c\}$$

est invariante pour le système (16).

Le système différentiel ordinaire associé à (16), obtenu en supprimant la viscosité, a deux points stationnaires, (0, 0, 0) et $(X_0, \gamma X_0/(1 + X_0), X_0)$, où X_0 est la solution positive d'une équation du second degré. Lorsque β est petit, ce point est instable, et on peut prouver l'existence de solutions homogènes en espaces et périodiques en temps (ce qui correspond à des configurations observées expérimentalement) ; par contre, pour β grand, le point stationnaire est stable et il n'y a pas de solution homogène périodique en temps. On écrit alors le système différentiel correspondant qui a une solution onde solitaire ; introduisant comme dans les exemples précédents la nouvelle variable $N = dX/d\xi$, $\xi = x - ct$, on obtient un nouveau système différentiel qui possède bien la nouvelle solution stationnaire :

$$(X_0, \gamma X_0/(1 + X_0), X_0, 0) ;$$

par contre, ce point est instable et on peut prouver qu'il existe dans son voisinage une

solution périodique, donc une onde solitaire périodique, qui, elle, sera stable pour le système d'évolution. On a ainsi montré que, pour β assez grand, on observe un phénomène périodique en espace temps.

Voici, en conclusion, quelques remarques générales.

1. Les équations de réaction-diffusion mettent en jeu la compétition entre des phénomènes non linéaires locaux et des phénomènes de diffusion en espace. Leur étude permet d'analyser l'apparition de solutions qui se propagent à cause de la diffusion, en conservant leur forme à cause de la non-linéarité. Cet aspect de compétition a aussi beaucoup été utilisé dans la dynamique des populations. On considère deux familles d'animaux (insectes, bactéries, etc.) vivant dans le même domaine Ω. On en déduit un système d'équations de réaction-diffusion, avec région invariante, dans lequel le facteur de diffusion dépend, à un changement d'échelle près, de la taille de Ω. Supposons que 0 est un attracteur stable ; on montre alors, si Ω est petit, que la solution tend toujours vers 0. Les prédateurs mangent toutes les proies, puis, n'ayant plus rien à manger, disparaissent. Par contre, si la diffusion est grande, on observe un phénomène d'oscillations en espace et en temps, comme dans la réaction de Bielousov-Zabotinski : les proies réussissent à s'échapper et à se reproduire avant d'être à nouveau atteintes par le prédateur et le phénomène se réitère.

2. La méthodologie a été de partir de l'analyse des équations différentielles ordinaires et d'en généraliser les outils classiques : régions invariantes, principes du maximum, points stables, etc. Nous en avons donné les aspects les plus simples. Le développement systématique des outils mathématiques dans ce domaine est dû à Conley et Smoller, une étape essentielle étant la généralisation de l'indice de Morse en dimension infinie.

3. Le terme de diffusion a été utilisé à propos de mouvements de population ou de densité de type brownien. Comme pour tous les problèmes qui introduisent des expressions de la forme :

$$\frac{\partial u}{\partial t} - \Delta u,$$

le rapport avec les probabilités est très étroit, et il existe aussi des analyses probabilistes de ce type de phénomènes.

CLAUDE BARDOS

Bibliographie

C. BARDOS, *Historique sommaire de l'équation de Korteweg et de Vries*, Institut de recherche sur l'enseignement des mathématiques, Villetaneuse, 1983 / R. COURANT & D. HILBERT, *Methods of Mathematical Physics*, Interscience, New York-Londres, 1962 / J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983 / G. WHITHAM, *Linear and Non Linear Waves*, Interscience-Wiley, New York, 1974.

DIFFÉRENTIELLES ÉQUATIONS

Les équations différentielles sont apparues historiquement tout au début du développement de l'analyse, en général à l'occasion de problèmes de mécanique ou de géométrie. Si, dans les premières investigations, l'on s'attachait surtout à en calculer les solutions au moyen de fonctions déjà connues, très vite ce point de vue s'affirma trop étroit ; c'est qu'en effet le problème fondamental de la théorie des équations différentielles est de déduire les propriétés des solutions d'une équation ou d'un système donné de la forme analytique de ceux-ci ; or, en général, les équations qui

résultent d'une investigation théorique en mathématiques ou en physique ne sont pas explicitement intégrables et constituent, bien souvent, la principale source pour la définition de nouvelles fonctions dont les propriétés peuvent être prévues par une analyse systématique de grandes classes d'équations ou de systèmes.

On développera, dans les quelques rubriques qui suivent, les méthodes propres à mettre en évidence l'existence de solutions sous des conditions appropriées et à en étudier les propriétés les plus fondamentales.



1. Les sytèmes différentiels linéaires dans le champ réel

On se propose d'étudier l'existence et les propriétés des solutions du système différentiel linéaire :

$$(1) \qquad dx_i/dt = \sum_{j=1}^{n} a_{ij}(t)x_j + b_i(t),$$

pour $i, j = 1, 2, ..., n$, où les fonctions $a_{ij}(t)$, $b_i(t)$ de la variable réelle t sont à valeurs réelles ou complexes. Introduisant la matrice $n \times n$, c'est-à-dire à n lignes et à n colonnes, $A(t) = (a_{ij}(t))$, et les vecteurs $x = (x_1, x_2, ..., x_n)$, $b = (b_1, b_2, ..., b_n)$, on peut écrire au lieu de (1) :

$$(2) \qquad dx/dt = A(t)x + b(t).$$

On notera que toute équation différentielle linéaire d'ordre n :

$$(3) \quad u^{(n)} + a_1(t)\, u^{(n-1)} + ... + a_n(t)\, u = b(t),$$

$u^{(j)}$ désignant la dérivée d'ordre j de la fonction $u(t)$ peut être ramenée à la forme (1) ou (2) au moyen de substitutions $x_1 = u$, $x_2 = u'$, ..., $x_n = u^{(n-1)}$, la matrice A et le vecteur b étant alors définis par :

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0. & 1 & \dots & 0 \\ \dots & \dots & \ddots & \ddots & \dots \\ -a_n & -a_{n-1} & \dots & \cdot & -a_1 \end{pmatrix}$$

$$b = (0, 0, ..., 0, b)$$

Existence des solutions

Un premier résultat fondamental est donné par le théorème suivant : Le système

$$(4) \qquad dx/dt = A(t)x,$$

$$(5) \qquad x(0) = c,$$

où $A(t)$ est une matrice $n \times n$ fonction continue de $t \in [0, t_0]$ et où c est un vecteur donné, a une solution unique $x(t)$ définie pour $t \in [0, t_0]$.

Il faut souligner qu'à l'équation (4) on a adjoint la condition initiale (5) ; on obtient ainsi un résultat d'existence et d'unicité.

On notera qu'au système (4), (5) on peut substituer l'équation intégrale équivalente :

$$(6) \qquad x(t) = c + \int_0^t A(\tau)x(\tau)\,d\tau,$$

qui se prête fort bien au calcul d'approximations successives inventé par Picard :

$$x_m(t) = c + \int_0^t A(\tau)x_{m-1}(\tau)\,d\tau,$$

avec $x_0 = c$ (cf. espaces MÉTRIQUES, chap. 7).

On établit la convergence de la suite $x_m(t)$ vers une fonction $x(t)$; on montre ensuite que $x(t)$ est solution de (4), (5) et qu'il y a unicité.

Le même type d'argument permet d'établir le théorème suivant : Le système

$$(7) \qquad dX/dt = A(t)\,X,$$

$$X(0) = D,$$

où $A(t)$ est une matrice $n \times n$ fonction continue de $t \in [0, t_0]$, et D une matrice $n \times n$ constante donnée, a une solution, matrice $n \times n$, $X(t)$, unique pour $t \in [0, t_0]$.

On réservera, dans la suite, la notation $X(t)$ à cette solution quand on prend pour D la matrice identité I, et l'on dira que $X(t)$ est la matrice résolvante. Le théorème de Jacobi montre que :

$$\det X(t) = \exp\left\{\int_0^t (\operatorname{tr} A)\, d\tau\right\},$$

où :

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}(\tau)\,;$$

ainsi la matrice $X(t)$ est toujours inversible.

Il est clair que la solution du système (4), (5) peut être représentée par $x(t) = X(t)c$.

En prenant pour c les éléments de la base de l'espace vectoriel $\mathbf{R}^n$ ou $\mathbf{C}^n$, on obtient n solutions de (4), qui sont les vecteurs dont les composantes sont inscrites successivement dans les colonnes de $X(t)$. Puisque $\det X(t) \neq 0$, les vecteurs sont indépendants quel que soit t. D'ailleurs, si l'on dispose de n solutions indépendantes à l'instant $t = 0$, elles le demeurent pour tout t : on dira que c'est un système fondamental de solutions. Enfin, il ne peut exister plus de n solutions indépendantes.

Dans le cas de l'équation différentielle (3), supposée homogène ($b(t) = 0$), les n solutions $u_1, u_2, ..., u_n$ sont indépendantes si, et seulement si, le déterminant de Wronski :

$$\det\begin{pmatrix} u_1 & u_2 & \dots & u_n \\ u'_1 & u'_2 & \dots & u'_n \\ u_1^{(n-1)} & u_2^{(n-1)} & \dots & u_n^{(n-1)} \end{pmatrix}$$

est $\neq 0$ pour tout t. Il suffit que cela soit vrai pour une valeur particulière de t.

L'équation linéaire non homogène

L'équation linéaire non homogène est l'équation :

(8) $\quad dx/dt = A(t)x + w(t), \quad x(0) = c.$

Toute solution de (8) où $A(t)$ et $w(t)$ sont respectivement une matrice $n \times n$ et un vecteur fonction continue donnée de $t \in [0, t_0]$ et c un vecteur constant peut être recherchée sous la forme : $x = X(t)\,y$, où $X(t)$ est la matrice résolvante de (7).

Il est aisé de voir que (8) conduit à :

$$dy/dt = X^{-1}(t)\, w(t),\ y(0) = c,$$

système qui a la solution unique :

$$y(t) = c + \int_0^t X^{-1}(\tau) w(\tau)\, d\tau,$$

d'où, pour (8), la solution unique :

$$x(t) = X(t)c + \int_0^t X(t)\, X^{-1}(\tau) w(\tau)\, d\tau.$$

Le cas des systèmes à coefficients constants

Si A est une matrice à éléments indépendants de t, la matrice résolvante $X(t)$ peut être représentée par la série convergente :

(9) $\quad X(t) = I + At/1! + ... + A^n t^n/n! ...$

On pourra introduire sur l'espace vectoriel des matrices carrées $n \times n$ la norme définie par :

$$\|A\| = \sum_{i,j=1}^{n} |a_{ij}|\,;$$

avec la topologie correspondante, on peut s'assurer que la série (9) converge uniformément par rapport à t sur tout intervalle fini, et satisfait (7). Par le théorème d'unicité on montre aisément que :

$$X(t)\, X(s) = X(s)\, X(t) = X(t + s),$$

ce qui, avec (9), suggère la définition :

$$e^{At} = X(t).$$

Il est connu que les puissances entières successives d'une matrice ne sont pas indépendantes ; si $f(\lambda)$ est le polynôme caractéristique $f(\lambda) = \det(A - \lambda I)$, on a $f(A) = 0$. Cela suggère que l'on peut donner à $X(t)$ une structure plus simple.

Pour tout nombre complexe λ, il existe un entier ν, le plus petit entier $\geqslant 0$ tel que :

$$(A - \lambda I)^{\nu+1}x = 0 \Rightarrow (A - \lambda I)^{\nu}x = 0\,;$$

$\nu = 0$ si, et seulement si, λ n'est pas valeur propre de la matrice A. Si λ est valeur propre, ν est au plus égale à l'ordre de multiplicité algébrique de λ. Soit $\lambda_1, ..., \lambda_k$, avec $k \leqslant n$, les valeurs propres distinctes de A, et soit :

$$\mathcal{M}_j = \{\, x : (A - \lambda_j I)^{\nu_j}x = 0 \,\}\,;$$

les sous-espaces $\mathcal{M}_j$ sont sans élément commun autre que 0 et leur somme directe est $\mathbf{C}^n$; on introduit les opérateurs de projection E_j par la formule $E_j x = x_j$, où $x = x_1 + x_2 + ... + x_k$, $x_j \in \mathcal{M}_j$ (décomposition spectrale). On établit que ces opérateurs permutent avec A et que :

$$(10)\quad X(t) = \sum_{j=1}^{k} \sum_{m=0}^{\nu_j - 1} \frac{(A - \lambda_j I)^m}{m\,!} t^m e^{\lambda_j t} E_{\lambda_j}$$

est la matrice résolvante de (7).

Cette formule fondamentale montre que les éléments de la matrice $X(t)$ sont des sommes de produits d'exponentielle $e^{\lambda_j t}$ par un polynôme en t dont le degré est inférieur à l'indice ν_j de λ_j, donc a fortiori à l'ordre de multiplicité algébrique de λ_j. On voit que les solutions de $dX/dt = Ax$ convergent toutes vers 0 quand $t \to +\infty$ si et seulement si les valeurs propres de la matrice A ont toutes leur partie réelle négative.

Le cas des systèmes à coefficients périodiques. La théorie de Floquet

Si $X(t)$ est la matrice résolvante de :

$$(11)\qquad dX/dt = P(t)X,$$

où $P(t)$ est une matrice $n \times n$ continue et périodique de période T par rapport à t, $X(t + T)$ vérifie (11) et en vertu du théorème d'unicité : $X(t + T) = X(t)\,X(T)$. Comme $X(T)$ est non singulière, il existe une matrice constante B telle que $X(T) = e^{BT}$. La matrice $Q(t) = X(t)e^{Bt}$ est périodique de période T et on obtient la représentation : $X(t) = Q(t)e^{-Bt}$. Les valeurs propres de la matrice B sont les coefficients caractéristiques de la matrice périodique $P(t)$; il faut et il suffit qu'ils soient tous à partie réelle négative pour que toute solution de $dx/dt = P(t)x$ tende vers 0 quand $t \to +\infty$.

Stabilité des solutions

Revenons au cas général $dx/dt = A(t)x$, où $A(t)$ est une matrice fonction continue de $t \in [0, +\infty]$; il est souvent utile de connaître certaines propriétés asymptotiques de la solution, par exemple, de savoir si elles demeurent bornées ou tendent vers 0 quand $t \to +\infty$. Pour conduire cette étude on utilise en général des méthodes de comparaison et, à cette fin, on peut introduire un concept de stabilité du genre suivant : les solutions de $dx/dt = A(t)x$ seront dites stables par rapport à une propriété $\mathfrak{P}$ et une classe $\mathfrak{F}$ de matrices $B(t)$ si les solutions de $dx/dt = (A(t) + B(t))x$ *ont toutes les propriétés* $\mathfrak{P}$ quel que soit $B(t) \in \mathfrak{F}$. On peut illustrer ce concept en citant le théorème suivant : Si les solutions $dx/dt = Ax$, où A est une matrice constante, sont bornées ou

tendent vers 0 quand $t \to +\infty$, alors il en est de même des solutions de :

(12) $$dx/dt = (\mathrm{A} + \mathrm{B}(t))x,$$

pourvu que :

$$\int_0^{+\infty} \| \mathrm{B}(t) \| dt < +\infty.$$

Si toutes les solutions de $dx/dt = \mathrm{A}x$ tendent vers 0 quand $t \to +\infty$, il en sera de même des solutions de (12) pourvu que $\| \mathrm{B}(t) \| \leqslant k$ pour t assez grand, k étant un nombre qui ne dépend que de A.

2. Les systèmes différentiels linéaires dans le champ complexe

On peut reprendre les problèmes discutés précédemment en supposant que les fonctions qui interviennent dans la définition du système (1) ou (2) sont des fonctions analytiques de la variable z dans un domaine Ω. On suppose d'abord que Ω est un domaine simplement connexe, c'est-à-dire un ensemble de points du plan complexe ouvert et connexe dont le complément par rapport au plan complexe muni du point à l'infini est connexe. On se propose de discuter le problème aux limites :

(13) $$dx/dz = \mathrm{A}(z)x, \quad x(z_0) = x_0,$$

avec $z_0 \in \Omega$ donné, x_0 vecteur de $\mathbf{C}^n$ donné, $\mathrm{A}(z)$ matrice $n \times n$ dont les éléments sont fonction holomorphe de z dans Ω.

On peut établir, en se servant de la méthode d'approximations successives, que le système (13) a une solution unique $x(z)$ holomorphe dans Ω. On peut aussi considérer le même problème pour le système matriciel :

(14) $$d\mathrm{X}/dz = \mathrm{A}(z)\mathrm{X}, \quad \mathrm{X}(z_0) = \mathrm{X}_0,$$

X_0 étant une matrice $n \times n$ donnée, et on parvient à une conclusion analogue, c'est-à-dire à l'existence d'une solution unique $\mathrm{X}(z)$ qui est une matrice $n \times n$ dont les éléments sont fonction holomorphe de z dans Ω.

Le théorème de Jacobi sous la forme :

$$\det \mathrm{X}(z) = \exp\left[\int_{z_0}^{z} \left(\sum_{i=1}^{n} a_{ii}(z) \right) dz\right] \det \mathrm{X}_0,$$

et les considérations antérieures sur les systèmes de solutions indépendantes développées dans le cas de variable réelle demeurent valables ici.

La structure des solutions dans le voisinage d'un point singulier

Une situation nouvelle apparaît si l'on suppose que la matrice $\mathrm{A}(z)$ possède des singularités ; plus précisément nous supposons que la singularité est en $z = 0$ et que $\mathrm{A}(z)$ est holomorphe dans le voisinage $0 < |z| < \mathrm{R}$; on précisera plus loin la nature de cette singularité : ce peut être un pôle (ce qui signifie que les éléments $a_{ij}(z)$ de $\mathrm{A}(z)$ ont tous au plus une singularité polaire en $z = 0$) ou une singularité essentielle (ce qui signifie que, parmi les éléments $a_{ij}(z)$, il en est un au moins qui possède, en $z = 0$, une singularité de cette nature).

D'après le résultat qui précède, on peut définir une solution de (14), $\mathrm{X}(z)$ matrice non singulière (on suppose $\det \mathrm{X}_0 \neq 0$), fonction holomorphe de z dans tout domaine simplement connexe où $\mathrm{A}(z)$ est holomorphe ; imaginons de choisir pour tel domaine un anneau de centre $z = 0$ dont la frontière est constituée des arcs de cercle $|z| = r$, $|z| = r'$, $0 < r < r' < \mathrm{R}$ et du segment joignant ces deux cercles et porté par le rayon qui passe par un certain point z, $r < |z| < r'$. Tournant autour de la singularité $z = 0$ de l'angle 2π, sur un

circuit contenu dans l'anneau, partant de z et y revenant, on pourra définir la matrice : $X^+(z) = X(ze^{2i\pi})$.

On s'assure aisément que $X^+(z)$ est solution locale de l'équation matricielle $dX/dz = A(z)X$, et, de là, qu'il existe une matrice constante non singulière U telle que $X^+(z) = X(z)\,U$.

Si on définit une matrice constante B telle $e^{2\pi i B} = U$ et si l'on pose $Q(z) = X(z)e^{-B \ln z}$, on voit que $Q(ze^{2i\pi}) = Q(z)$, c'est-à-dire que $Q(z)$ est une fonction uniforme ; c'est d'ailleurs, une fonction holomorphe de z dans le voisinage de $z = 0$, sauf peut-être en $z = 0$. On obtient alors la représentation fondamentale :

$$X(z) = Q(z)e^{B \ln z},$$

où $Q(z)$ est une matrice inversible fonction holomorphe de z dans $0 < |z| < R$. En outre, si $\lambda_1, ..., \lambda_k$ sont les valeurs propres distinctes de B, $\mathcal{M}_1, ..., \mathcal{M}_k$, les sous-espaces de la décomposition spectrale, et E_j les projecteurs associés, on peut écrire la formule précédente :

$$(15) \quad X(z) = Q(z) \times \sum_j \sum_{m=0}^{\nu_j - 1} \frac{(B - \lambda_j I)^m}{m!} z^{\lambda_j} (\ln z)^m E_j.$$

Pour achever de préciser la structure de $X(z)$ dans le voisinage de $z = 0$, il faut savoir quelle est la nature du point $z = 0$ pour la matrice $Q(z)$. Or, celle-ci étant une fonction holomorphe uniforme dans le voisinage de $z = 0$, deux cas seuls sont possibles : elle est holomorphe en $z = 0$, ou bien elle a une singularité en ce point, qui ne peut être qu'un pôle ou une singularité essentielle.

On dit que le système $dX/dz = A(z)X$ est du type de Fuchs en $z = 0$ si la matrice $Q(z)$ a au plus un pôle en ce point. Le critère de régularité de Birkhoff exprime une condition nécessaire et suffisante pour qu'il en soit ainsi : $\|X(z)\| \times |z|^\alpha$ doit être borné dans $0 < |z| < R$ pour une certaine valeur de α. Ces considérations s'appliquent pour la description des solutions de l'équation différentielle linéaire d'ordre n :

$$(16) \quad u^{(n)} + a_1(z)u^{(n-1)} + ... + a_n(z)u = 0,$$

où les $a_j(z)$ sont fonction holomorphe de z dans $0 < |z| < R$.

Cependant, dans ce dernier cas, on peut énoncer le théorème (Fuchs) : L'équation (16) est du type de Fuchs en $z = 0$ si et seulement si $a_j(z)$ a, en $z = 0$, un pôle d'ordre j au plus, avec $1 \leqslant j \leqslant n$.

Le cas des équations différentielles linéaires du second ordre

D'après ce qui précède, l'équation différentielle du second ordre la plus générale qui a en $z = 0$ une singularité du type de Fuchs peut s'écrire :

$$(17) \quad d^2u/dz^2 + p(z)\,du/dz + q(z)u = 0,$$

avec :

$$(18) \quad \begin{aligned} zp(z) &= p_0 + p_1 z + ... + p_n z^n + ... \\ z^2 q(z) &= q_0 + q_1 z + ... + q_n z^n + ..., \end{aligned}$$

les séries $\Sigma p_n z^n$, $\Sigma q_n z^n$ étant convergentes dans le disque $|z| < R$. On peut chercher formellement des solutions de (17) en posant :

$$(19) \qquad u = z^\alpha \Big(1 + \sum_{n \geqslant 1} c_n z^n\Big),$$

α et les c_n étant des coefficients à déterminer.

Par substitution dans (17), compte tenu de (18), on peut établir que les coefficients c_n sont déterminés de proche en proche de manière unique pourvu que α soit pris égal à l'une des racines de l'équation :

$$(20) \quad F(\alpha) \equiv \alpha^2 + (p_0 - 1)\alpha + q_0 = 0$$

et que $F(\alpha + n)$ ne soit jamais nul pour n entier > 0.

Par conséquent, si la différence des racines $\alpha_1 - \alpha_2$ de l'équation déterminante (20) est non nulle et non égale à un entier, on pourra ainsi obtenir deux développements du type (19) satisfaisant l'équation (17) et, dans tous les cas, on en aura au moins un.

On peut établir la convergence de ces développements par une méthode de majorante et l'on voit qu'on aura toujours une solution du type (19) :

$$u_1(z) = z^{\alpha_1}\Big(1 + \sum_{n \geqslant 1} c_n z^n\Big),$$

avec $F(\alpha_1) = 0$, $F(\alpha_1 + n) \neq 0$, n entier > 0, et, éventuellement, une autre solution indépendante du même type :

$$u_2(z) = z^{\alpha_2}\Big(1 + \sum_{n \geqslant 1} d_n z^n\Big),$$

si $\alpha_1 - \alpha_2$ n'est pas entier. Dans ce cas on dira que l'équation est du premier type de Fuchs en $z = 0$.

Si maintenant $\alpha_1 - \alpha_2 = s$, entier $\geqslant 0$, on peut faire dans (17) le changement de fonction inconnue : $u = u_1(z)\zeta(z)$ et calculer aisément $\zeta(s)$, ce qui conduit aux résultats suivants :

a) s non nul :

$$u_2(z) = \mathrm{A}u_1(z) + \mathrm{B}[gu_1(z)\ln z + \tilde{u}(z)],$$

avec : $$\tilde{u}(z) = z^{\alpha_2}\Big(1 + \sum_{1}^{\infty} h_n z^n\Big);$$

b) s nul :

$$u_2(z) = \mathrm{A}u_1(z) + \mathrm{B}\Big[u_1(z)\ln z + z^{\alpha_2}\sum_{n=1}^{\infty} h_n z_n\Big],$$

A et B désignant des constantes arbitraires.

On dira alors que l'équation est du second type de Fuchs en $z = 0$. Pour des raisons de commodité, nous avons discuté de la structure des solutions dans le voisinage du point singulier $z = 0$; mais, par le moyen d'une translation ou de la transformation $z \mapsto 1/z$, on pourra toujours ramener à l'origine une singularité quelconque du plan complexe, qu'elle soit à distance finie ou infinie.

Les équations différentielles de la physique mathématique

Les équations différentielles linéaires du second ordre dont les coefficients sont fonction analytique de z, ayant pour seuls points singuliers z_1, z_2, z_3, z_4, et l'infini, ceux-ci étant du type de Fuchs, avec, pour exposants, α_j, β_j en z_j, μ_1, μ_2 à l'infini (racines de l'équation déterminante), sont nécessairement du type :

$$\text{(21)}\quad \frac{d^2u}{dz^2} + \Big(\sum_{j=1}^{4}\frac{1-\alpha_j-\beta_j}{z-z_j}\Big)\frac{du}{dz} + \left(\sum_{j=1}^{4}\frac{\alpha_j\beta_j}{(z-z_j)^2} + \frac{\mathrm{A}z^2 + 2\,\mathrm{B}z + \mathrm{C}}{\prod_{j=1}^{4}(z-z_j)}\right)u = 0,$$

où A est tel que μ_1 et μ_2 sont racines de :

$$\mu^2 + \mu\Big(\sum_{j=1}^{4}(\alpha_j + \beta_j) - 3\Big) + \sum_{j=1}^{4}\alpha_j\beta_j + \mathrm{A} = 0,$$

et B et C des constantes arbitraires. On notera la condition nécessaire :

$$\sum_{j=1}^{4}(\alpha_j + \beta_j) + \mu_1 + \mu_2 = 3.$$

Il a été établi par Klein et Bocher que de nombreuses équations différentielles qui apparaissent dans certaines branches

de la physique mathématique sont des formes confluentes de l'équation spéciale de type (21) obtenue quand la différence des exposants à chaque singularité est égale à 1/2 (équation de Lamé généralisée). En général, par confluence de deux singularités, on obtient encore un point singulier du type de Fuchs, mais cela n'est plus vrai quand on envisage la confluence de trois singularités. En bref, on peut, par des confluences convenables des cinq singularités qui sont à notre disposition, obtenir six types d'équations qu'on peut classer suivant :

a) le nombre de leurs points singuliers du premier type de Fuchs avec des exposants dont la différence vaut 1/2,

b) le nombre des autres points singuliers du type de Fuchs,

c) le nombre de points singuliers qui ne sont pas du type de Fuchs :

	(*a*)	(*b*)	(*c*)	
I	3	1	0	Lamé
II	2	0	1	Mathieu
III	1	2	0	Legendre
IV	0	1	1	Bessel
V	1	0	1	Weber-Hermite
VI	0	0	1	Stokes

Donnons un exemple : $z_1 = z_2 = 0$, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, on fait tendre z_3 et z_4 vers l'infini. On obtient l'équation :

$$z^2 d^2u/dz^2 + z\,du/dz + (z - n^2)u/4 = 0,$$

qui, par le changement de variable $z \mapsto z^2$, devient l'équation de Bessel :

$$z^2 d^2u/dz^2 + z\,du/dz + (z^2 - n^2)u = 0.$$

De façon analogue, l'équation différentielle linéaire du second ordre la plus générale à coefficients qui soient des fonctions analytiques de z, ayant pour seuls points singuliers a, b, c, ceux-ci étant du type de Fuchs, avec, pour exposants, $\alpha, \alpha', \beta, \beta', \gamma, \gamma'$, peut s'écrire (Riemann) :

$$(22)\quad \frac{d^2u}{dz^2} + \left[\frac{1-\alpha-\alpha'}{z-a} + \frac{1-\beta-\beta'}{z-b} + \frac{1-\gamma-\gamma'}{z-c}\right]\frac{du}{dz} + \left[\frac{\alpha\alpha'(a-b)(a-c)}{z-a} + \frac{\beta\beta'(b-a)(b-c)}{z-b} + \frac{\gamma\gamma'(c-a)(c-b)}{z-c}\right] \times \frac{u}{(z-a)(z-b)(z-c)} = 0.$$

On exprime que u est une solution de (22) au moyen de la notation :

$$u = \mathbf{P}\begin{Bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{Bmatrix}.$$

Par exemple, l'équation hypergéométrique est du type de (22) :

$$z(1-z)\,d^2u/dz^2 + [\gamma - (\alpha+\beta+1)z]\,du/dz - \alpha\beta u = 0,$$

$$u = \mathbf{P}\begin{Bmatrix} 0 & \infty & 1 & \\ 0 & \alpha & 0 & z \\ 1-\gamma & \beta & \gamma-\alpha-\beta & \end{Bmatrix}.$$

Dans le cas général de l'équation de Riemann (22), les deux transformations suivantes qu'on peut vérifier par un calcul direct sont de grande importance :

$$\left(\frac{z-a}{z-b}\right)^k \times \left(\frac{z-c}{z-b}\right)^l \times \mathbf{P}\begin{Bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{Bmatrix} = \mathbf{P}\begin{Bmatrix} a & b & c & \\ \alpha+k & \beta-k-l & \gamma+l & z \\ \alpha'+k & \beta'-k-l & \gamma'+l & \end{Bmatrix}$$

$$\mathbf{P}\begin{Bmatrix} a & b & c & \\ \alpha & \beta & \gamma & z \\ \alpha' & \beta' & \gamma' & \end{Bmatrix} = \mathbf{P}\begin{Bmatrix} a_1 & b_1 & c_1 & \\ \alpha & \beta & \gamma & z_1 \\ \alpha' & \beta' & \gamma' & \end{Bmatrix}$$

où $(z_1, a_1, b_1, c_1) = (z, a, b, c)$, égalité des rapports anharmoniques.

3. Le problème de Sturm-Liouville

Position du problème

Considérons l'équation différentielle linéaire d'ordre n :

(23) $$L(u) \equiv p_0(t)\, d^n u/dt^n + p_1(t)\, d^{n-1}u/dt^{n-1} + \ldots + p_n(t)\, u = r(t),$$

où $p_0(t)$, ..., $p_n(t)$, $r(t)$ sont des fonctions continues de la variable réelle t dans l'intervalle $t \in [a, b]$, et $p_0(t) \neq 0$ pour $t \in [a, b]$.

Imaginons un ensemble de m conditions aux limites du type :

(24) $$U_i(u) \equiv \alpha_i^0 u(a) + \alpha_i^1 u'(a) + \ldots + \alpha_i^{n-1} u^{(n-1)}(a) + \beta_i^0 u(b) + \beta_i^1 u'(b) + \ldots + \beta_i^{n-1} u^{(n-1)}(b) = \gamma_i, \quad 1 \leqslant i \leqslant m,$$

où les α_i^j, β_i^j, γ_i sont des nombres donnés, les formes linéaires (24) étant supposées linéairement indépendantes par rapport aux $2n$ variables $u(a)$, $u'(a)$, ..., $u^{(n-1)}(a)$, $u(b)$, ..., $u^{(n-1)}(b)$, ce qui implique $m \leqslant 2n$. On se propose de rechercher dans la classe des fonctions différentiables jusqu'à l'ordre n pour $t \in [a, b]$ s'il existe des solutions du système :

(25) $$Lu = r(t) \qquad U_i(u) = \gamma_i, \quad 1 \leqslant i \leqslant m.$$

Il semble commode de commencer cette étude par la discussion du problème homogène :

(26) $$Lu = 0 \qquad U_i(u) = 0, \quad 1 \leqslant i \leqslant m.$$

Pour résoudre (26), il suffit une fois construit un système fondamental de solutions de l'équation $L(u) = 0$, soit $u_1(t)$, $u_2(t)$, ..., $u_n(t)$, de rechercher la solution du système (26) sous la forme :

$$u = c_1 u_1 + c_2 u_2 + \ldots + c_n u_n,$$

les n coefficients c_1, ..., c_n devant alors être calculés par le moyen des m équations linéaires à n inconnues :

(27) $$\sum_{j=1}^{n} c_j U_i(u_j) = 0, \quad 1 \leqslant i \leqslant m.$$

Si le rang de la matrice $U_i(u_j)$ est égal à $p \leqslant n$, il existe $n - p$ vecteurs $\{ c_1, c_2 \ldots, c_n \}$ indépendants vérifiant (27). Si le rang de cette matrice est égal à n, ce qui implique $m \geqslant n$, le système (27) n'a pas de solution hormis la solution banale $u = 0$.

Des considérations analogues peuvent être développées pour le système (25).

Le système adjoint

Supposant que les coefficients $p_j(t)$ ont des dérivées par rapport à t jusqu'à l'ordre $n - j$, introduisons l'opérateur :

(28) $$\tilde{L}(u) = (-1)^n d^n(p_0 v)/dt^n + (-1)^{n-1} d^{n-1}(p_1 v)/dt^{n-1} + \ldots + (-1)\, d(p_{n-1} v)/dt + p_n v$$

grâce auquel on obtient l'identité de Lagrange :

$$vL(u) - u\tilde{L}(v) = (d/dt)\, P(u, v),$$

où $P(u, v)$ est une forme bilinéaire homogène par rapport aux deux ensembles de variables u, u', ..., $u^{(n-1)}$, et v, v', ..., $v^{(n-1)}$. Par intégration on obtient :

(29) $$\int_a^b [v\, L(u) - u\, \tilde{L}(v)]dt = P(u, v)\,|_{t=b} - P(u, v)\,|_{t=a}.$$

Complétons le système des m formes linéaires données $U_i(u)$ par $2n - m$ formes linéaires des variables $u(a)$, ..., $u^{(n-1)}(a)$, $u(b)$, ..., $u^{(n-1)}(b)$, de telle sorte que le système des $2n$ formes ainsi obtenu soit encore indépendant. On peut montrer qu'il existe alors un système de $2n$ formes linéaires indépendantes des $2n$ variables

$v(a)$, $v'(a)$, ..., $v^{(n-1)}(a)$, $v(b)$, $v^{(n-1)}(b)$, soit $V_1(v)$ ... $V_{2n}(v)$ tel que l'on ait l'identité :

$$(30)\quad P(u, v)|_b - P(u, v)|_a = U_1 V_{2n} + U_2 V_{2n-1} + ... + U_{2n} V_1.$$

Supposant U_1, U_2, ..., U_m fixées, remplaçons les formes U'_{m+1}, ..., U'_{2n} des mêmes variables constituant avec U_1, ..., U_m un système indépendant. Dans la formule (30), les formes V_1, ..., V_{2n-m} seront alors changées en des formes V'_1, ..., V'_{2n-m}, mais ces deux derniers systèmes de $2n - m$ formes sont équivalents. Le système V_1 ... V_{2n-m} ne dépend donc en tant que système de formes linéaires, que des m formes données U_1, ..., U_m, ce qui justifie la définition suivante : Le système :

$$(31)\qquad \tilde{L}(v) = 0 \qquad V_i(v) = 0,$$

avec $i = 1, 2, ..., 2n - m$, est dit *adjoint* au système (26).

Le cas des systèmes différentiels autoadjoints du second ordre

Un cas très important pour les applications est celui où :

$$L(u) = p_0(t)\, d^2u/dt^2 + p_1(t)\, du/dt + p_2(t)u.$$

Une condition nécessaire et suffisante pour que $L(u) = \tilde{L}(u)$ est que $p_0' = p_1$, et dans ce cas on peut écrire :

$$L(u) = (d/dt)(p_0\, du/dt) + p_2 u.$$

Toutefois, si $L(u)$ n'est pas autoadjoint, on peut le rendre tel par multiplication par un facteur convenable :

$$\frac{1}{p_0}\exp\left\{\int\frac{p_1}{p_0}dt\right\}L(u) = \frac{d}{dt}\left[\exp\left\{\int\frac{p_1}{p_0}dt\right\}\frac{du}{dt}\right] + \frac{p_2}{p_0}\exp\left\{\int\frac{p_1}{p_0}dt\right\}u.$$

Puisque, ainsi, toute équation du second ordre peut être réduite à sa forme autoadjointe, on peut se borner à l'étude d'équation du type :

$$(d/dt)(k\, du/dt) - gu = r,$$

connue sous le nom d'équation de Sturm-Liouville.

Étant donné le problème aux limites :

$$L(u) \equiv (d/dt)(k\, du/dt) - gu = 0, \qquad k(t) \neq 0, \quad t \in [a, b],$$
$$U_1(u) \equiv \alpha_1 u(a) + \alpha_2 u(b) + \alpha_3 u'(a) + \alpha_4 u'(b) = 0,$$
$$U_2(u) = \beta_1 u(a) + \beta_2 u(b) + \beta_3 u'(a) + \beta_4 u'(b) = 0,$$

on peut construire le système adjoint :

$$L(v) \equiv (d/dt)(k\, dv/dt) - gv = 0 \qquad V_1(v) = 0 \qquad V_2(v) = 0,$$

et chercher sous quelles conditions il est autoadjoint, ce qui revient à exprimer que le système des formes $U_1(u)$, $U_2(u)$ est équivalent au système $V_1(u)$, $V_2(u)$. On montre qu'il en est ainsi si et seulement si :

$$d_{24}k(a) = d_{13}k(b), \text{ avec } d_{ij} = \alpha_i\beta_j - \alpha_j\beta_i.$$

Exemple :

$$U_1(u) = u'(a) - hu(a), \; U_2(u) = u'(b) + Hu(b).$$

Le problème de Sturm-Liouville peut être posé comme suit : étant donné l'équation

$$(32)\quad L(u) = (d/dt)(k(t)\, du/dt) - g(t)u = \lambda r(t)u,$$

où $k(t)$, $g(t)$, $r(t)$ sont fonctions continues de t, $k(t) \neq 0$ pour $t \in [a, b]$ et les conditions aux limites :

$$(33)\qquad U_1(u) = 0, \quad U_2(u) = 0,$$

trouver les valeurs du paramètre λ pour lesquelles le système (32), (33) a des solutions régulières dans $t \in [a, b]$ et construire ces solutions.

La fonction de Green

Revenons au cas général de l'équation d'ordre n :

$$(34)\quad L(u) = p_0(t)d^n u/dt^n + p_1(t)d^{n-1}u/dt^{n-1} + \dots + p_{n-1}(t)\,du/dt + p_n(t)\,u = 0,$$

$$(35)\quad U_i(u) = 0,\quad i = 1, 2, \dots, n,\quad p_0(t) \neq 0,\quad t \in [a, b].$$

On suppose le système (34), (35) incompatible, c'est-à-dire qu'il n'existe pas de solution régulière autre que la solution nulle. On peut montrer alors qu'il existe une fonction $G(t, t')$, dite fonction de Green, et une seule, satisfaisant aux propriétés suivantes :

1. Elle est continue et possède des dérivées continues jusqu'à l'ordre $n-2$ inclus par rapport à $t \in [a, b]$;

2. Elle est telle que sa dérivée d'ordre $n-1$ par rapport à t existe et est continue pour $t \in [a, b]$ sauf en $t = t'$; en ce point il y a discontinuité de cette dérivée, le saut valant $1/p_0(t')$;

3. Elle satisfait au système (35) et à l'équation (34) relativement à la variable t sauf en $t = t'$.

Si $H(t, t')$ est la fonction de Green associée au système adjoint de (34), (35), on peut montrer que $G(t, t') = H(t', t)$; en particulier, dans le cas d'un système autoadjoint, $G(t, t') = G(t', t)$ et la fonction de Green est symétrique par rapport à ses deux variables.

Si le système (34), (35) est incompatible, le système non homogène :

$$(36)\quad \begin{aligned} &L(u) = r(t) \\ &U_i(u) = 0,\quad 1 \leqslant i \leqslant n, \end{aligned}$$

a une solution unique, qu'on construit aisément avec la fonction de Green :

$$u(t) = \int_a^b G(t, t')r(t')\,dt'.$$

Le problème aux limites de Sturm-Liouville (32), (33) est donc, sous l'hypothèse que le système homogène :

$$(37)\quad L(u) = 0,\quad U_1(u) = 0,\quad U_2(u) = 0$$

soit incompatible, équivalent à l'équation intégrale :

$$(38)\quad u(t) = \lambda \int_a^b G(t, t')r(t')u(t')\,dt'.$$

Exemple de fonction de Green : soit

$$L \equiv d^2/dt^2,\quad [a, b] = [0, 1],$$
$$U_1(u) = u(0),\quad U_2(u) = u(1),$$

on trouve :

$$\begin{aligned} G(t, t') &= (1 - t')t,\quad 0 \leqslant t < t' \leqslant 1, \\ &= (1 - t)t',\quad 0 \leqslant t' < t \leqslant 1. \end{aligned}$$

Les fonctions propres et la théorie de Hilbert-Schmidt

Si le système (37) est autoadjoint, la fonction de Green $G(t, t')$ est symétrique, et si l'on fait l'hypothèse $r(t) > 0$, quel que soit $t \in [a, b]$, posant :

$$y(t) = [r(t)]^{1/2} u(t),$$
$$K(t, t') = G(t, t')[r(t)\,r(t')]^{1/2},$$

l'équation (38) peut s'écrire :

$$(39)\quad y(t) = \lambda \int_a^b K(t, t')y(t')dt',$$

où le noyau $K(t, t')$ est une fonction symétrique continue des deux variables t, t' dans le carré $t \in [a, b]$, $t' \in [a, b]$. On est ainsi conduit à une équation intégrale de Fredholm à noyau symétrique.

Les valeurs λ pour lesquelles (39) a des solutions non nulles sont les valeurs propres, les solutions correspondantes étant les fonctions propres. Il est commode, pour l'énoncé des résultats, d'introduire ici la notion de produit scalaire de deux

fonctions $f(t)$, $g(t)$ continues dans $t \in [a, b]$; on notera ce produit :

$$(f|g) = \int_a^b f(t)\,\bar{g}(t)\,dt,$$

où $\bar{g}$ est la valeur complexe conjuguée de g ; f sera dit orthogonal à g si $(f|g) = 0$.

On établit les résultats suivants (cf. espace de HILBERT, théorie SPECTRALE) :

1. Toutes les valeurs propres de (39) sont réelles ; deux fonctions propres qui correspondent à deux valeurs propres distinctes sont orthogonales.

2. L'ensemble des valeurs propres est dénombrable et n'a pas d'autre point limite que l'infini, éventuellement. La multiplicité de chaque valeur propre est finie, ce qui signifie que le nombre des fonctions propres linéairement indépendantes correspondant à toute valeur propre λ est fini.

On peut donc écrire la suite des valeurs propres :

$$|\lambda_1| \leqslant |\lambda_2| \dots \leqslant |\lambda_j| \leqslant \dots,\quad \lim_{j\to\infty} |\lambda_j| = \infty,$$

et des fonctions propres correspondantes :

$$\varphi_1, \varphi_2, \dots, \varphi_j, \dots,$$

si la suite est infinie, et l'on peut, grâce au procédé d'orthogonalisation de Schmidt, faire en sorte que le système des fonctions propres φ_j soit orthonormé, c'est-à-dire :

$$(\varphi_i|\varphi_j) = \delta_{ij} = 1 \quad \text{si } i = j$$
$$= 0 \quad \text{si } i \neq j.$$

3. Soit $f(t)$ une fonction continue de $t \in [a, b]$. Les nombres $(f|\varphi_j), j = 1, 2, \dots$ sont appelés les coefficients de Fourier de f.

Avec :

$$\|f\|^2 = \int_a^b |f|^2\,dt,$$

on a l'inégalité de Bessel :

$$\|f\|^2 \geqslant \sum_{j=1}^{\infty} |(f|\varphi_j)|^2.$$

4. *Théorème de Hilbert-Schmidt.* Soit $f(t)$ une fonction continue de $t \in [a, b]$ et :

$$\text{(40)} \qquad \mathrm{K}f = \int_a^b \mathrm{K}(t, t')f(t')\,dt'.$$

Le développement de Fourier de $\mathrm{K}f$ par rapport au système orthonormé de toutes les fonctions propres φ_j, qui s'écrit :

$$\sum_{j=1}^{\infty} (\mathrm{K}f|\varphi_j)\varphi_j(t),$$

converge absolument vers $\mathrm{K}f$ et uniformément par rapport à $t \in [a, b]$.

Exemple : soit $y(t)$ une fonction continue pourvu de dérivées première et seconde continues pour $t \in [0, 1]$ et telle que $y(0) = y(1) = 0$. Alors $y(t)$ satisfait à $y''(t) = r(t)$, $r(t)$ continue.

Par suite, $y(t)$ peut être développée en série de Fourier :

$$y(t) = \sum_1^{\infty} \alpha_n \sqrt{2} \sin(n\pi t),$$

avec :

$$\alpha_n = \int_0^1 y(t)\,\sqrt{2}\sin(n\pi t)\,dt,$$

la convergence étant uniforme sur l'intervalle [0, 1]. On aura reconnu que $\sqrt{2}\sin(n\pi t)$, $n = 1, 2$, est le système orthogonal des fonctions propres du système $y'' = \lambda y$, $y(0) = y(1) = 0$.

La présentation donnée ici, qu'on a voulu très élémentaire, a été le point de départ de généralisations très importantes et très fructueuses qui sont au cœur des mathématiques contemporaines (espace de Hilbert, opérateurs linéaires, théorie spec-

trale, etc.). D'autre part, nous avons supposé l'intervalle $[a, b]$ fini et les fonctions k, g, r continues sur l'intervalle fermé, $k(t)$ ne s'annulant jamais. Si l'une de ces conditions est omise, les conclusions indiquées ne sont plus valables et de nouveaux problèmes apparaissent dans le détail desquels il est impossible d'entrer ici. Ces problèmes sont connus sous le nom de problèmes aux limites singuliers et les résultats fondamentaux obtenus dans ce domaine sont associés aux noms de H. Weyl, M. H. Stone, E. C. Tichmarsh, K. Kodaira, en particulier le théorème fondamental de développement qui généralise celui de Hilbert-Schmidt (spectre continu) et qui permet de donner une présentation unitaire de la théorie du développement des fonctions en série de Fourier, série de fonctions d'Hermite, fonctions de Bessel, intégrale de Fourier, etc.

Il n'est pas sans intérêt, pour conclure, de donner quelques brèves indications sur le développement historique de la question c'est-à-dire la méthode de Sturm-Liouville proprement dite.

Associons à l'équation (32) les conditions aux limites :

$$\alpha_1 u(a) + \alpha_2 u'(a) = 0\,; \quad \beta_1 u(b) + \beta_2 u'(b) = 0\,; \tag{41}$$

on peut définir une solution de (32) satisfaisant à la première condition (41), $u = u(t, \lambda)$ pour $t \in [a, b]$. Substituant dans la deuxième condition (41), on est conduit à résoudre l'équation :

$$F(\lambda) = \beta_1 u(b, \lambda) + \beta_2 u'(b, \lambda) = 0.$$

On conçoit que, dans ce but, l'étude des zéros de $u(t, \lambda)$ dans l'intervalle $[a, b]$ et la manière dont ils évoluent quand λ varie puissent être de quelque utilité. Pour cette étude, on dispose de théorèmes de comparaison (Sturm), qui, grosso modo, indiquent comment varient ces zéros quand on remplace dans l'équation (32) $k(t)$ et $p(t) = \lambda\, r(t) + g(t)$ par des fonctions $k_1(t)$ et $p_1(t)$ telles que $k_1 \geqslant k, p_1 \geqslant p$ quel que soit $t \in [a, b]$.

4. Les systèmes différentiels non linéaires

Les systèmes différentiels non linéaires dans le champ réel

On considère le système différentiel :

$$dx/dt = f(x, t), \tag{42}$$

où $x \in \mathbf{R}^n$, $f(x, t)$ une fonction à valeurs dans $\mathbf{R}^n$, t une variable réelle. On suppose $f(x, t)$ définie et continue dans l'ensemble $\bar{G} \times [t_0, t_0 + T]$, où G est un ensemble ouvert et borné dans $\mathbf{R}^n$.

Avec x_0 donné dans G, on se propose de discuter le problème aux limites :

$$dx/dt = f(x, t), \quad x(t_0) = x_0. \tag{43}$$

On peut imaginer le procédé constructif suivant : soit $t_0 < t_1 < t_2 \ldots < t_p < \ldots$, une suite finie de valeurs de t et :

$$\begin{aligned} x_1 &= x_0 + (t_1 - t_0) f(x_0, t_0), \\ x_2 &= x_1 + (t_2 - t_1) f(x_1, t_1), \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ x_p &= x_{p-1} + (t_p - t_{p-1}) f(x_{p-1}, t_{p-1}), \end{aligned}$$

et introduisons la fonction $\tilde{x}(t)$ à valeurs dans $\mathbf{R}^n$ définie par :

$$\begin{aligned} \tilde{x}(t) &= x_0 + (t - t_0) f(x_0, t_0), \quad t_0 \leqslant t \leqslant t_1 \\ &= x_1 + (t - t_1) f(x_1, t_1), \quad t_1 \leqslant t \leqslant t_2. \end{aligned}$$

La fonction $\tilde{x}(t)$ est évidemment continue, sa représentation dans $\mathbf{R}^n$ étant une ligne polygonale. On peut espérer, si les intervalles $t_{j+1} - t_j$ ne sont pas trop grands ou mieux tendent vers 0, que $x(t)$

tendra d'une certaine façon vers une fonction $x(t)$ solution de (43).

Tel est le principe de la méthode des différences finies dont les applications débordent largement le cadre de la théorie des équations différentielles. On peut établir ainsi le théorème : Si $f(x, t)$ est continue dans

$$\overline{G} \times [t_0, t_0 + T],$$

où G est un ensemble ouvert et borné de $\mathbf{R}^n$, alors pour tout $x_0 \in G$, il existe une fonction vectorielle $x(t)$ solution de (43) dans l'intervalle $[t, t + h]$, où :

$$h = \inf(T, d/M), \quad M = \sup_{\substack{x \in G \\ t \in [t_0, t_0 + T]}} \|f\|,$$

d étant la distance de x_0 à la frontière de G. On pourra prendre pour définition de la norme de f, ou de x, $\|f\|$, $\|x\|$, la somme des modules des composantes ; on a :

$$d = \sup \delta, \ \delta \text{ tel que } \|x - x_0\| < \delta \Rightarrow x \in G.$$

On notera que cet énoncé exprime seulement l'existence d'une solution ; si l'on fait l'hypothèse additionnelle :

$$(44) \quad \|f(x', t) - f(x'', t)\| \leqslant k \|x' - x''\|,$$

quels que soient $(x', x'') \in G$, $t \in [t_0, t_0 + T]$, où k est une constante (condition de Lipschitz), alors on peut établir qu'il y a unicité.

D'autre part, les propriétés de la solution à l'égard des conditions initiales peuvent être appréciées par le résultat suivant : si l'on suppose que $f(x, t)$ a des dérivées partielles premières par rapport aux composantes de x continues dans $\|x - x_0\| \leqslant \delta$, $t \in [t_0, t_0 + T]$, alors la solution $x(t)$ du système (43), qui existe et est unique dans $t \in [t_0, t_0 + h]$, a des dérivées partielles premières continues par rapport aux composantes de x_0.

Une autre méthode d'approche est celle des approximations successives, que nous avons déjà rencontrée. On fait l'hypothèse (44) et l'on définit la suite $x_1(t)$, ..., $x_m(t)$ par :

$$dx_1/dt = f(x_0, t), \qquad x_1(t_0) = x_0$$
$$\dots\dots\dots\dots\dots\dots$$
$$dx_{m+1}/dt = f(x_m(t), t), \quad x_m(t_0) = x_0$$
$$\dots\dots\dots\dots\dots\dots$$

équivalente à :

$$x_1(t) = x_0 + \int_{t_0}^{t} f(x_0, \tau)\, d\tau$$
$$\dots\dots\dots\dots\dots\dots$$
$$x_{m+1}(t) = x_0 + \int_{t_0}^{t} f(x_m(\tau), \tau)\, d\tau$$
$$\dots\dots\dots\dots\dots\dots$$

On montre aisément que cette suite peut être construite dans $t \in [t_0, t_0 + h]$ et converge uniformément dans cet intervalle vers une fonction $x(t)$ solution de (43) ; le même type d'argument permet d'établir l'unicité (cf. chap. 7 *Intégration numérique des équations différentielles* pour l'aspect numérique).

Les systèmes différentiels non linéaires dans le champ complexe

On peut développer des théorèmes d'existence locale assez voisins de ceux qui sont décrits dans le cas précédent. Mais l'étude des solutions d'un point de vue global est très instructive et amène à des distinctions intéressantes lorsqu'on discute de leurs singularités.

Celles-ci sont, en général, de deux sortes : les singularités fixes qu'on peut prévoir a priori d'après la nature du système différentiel donné et celles qui sont mobiles, c'est-à-dire dépendent des conditions initiales. Pour simplifier, on se bornera, dans l'exposé qui suit, au seul cas des équations différentielles non linéaires.

Dans le cas d'une équation du premier ordre : $du/dz = f(z, u)$, f fonction analy-

tique de z et rationnelle en u, on démontre qu'il ne peut pas exister de singularité essentielle mobile, quoique des singularités mobiles plus simples puissent apparaître (pôle ou point de branchement). Exemple :

$$du/dz + z/u = 0, \quad u = \sqrt{z_0^2 + u_0^2 - z^2},$$
$$u(z_0) = u_0.$$

Mais, dans le cas d'équation du second ordre, ou d'ordre plus élevé, des singularités essentielles mobiles peuvent apparaître comme l'indique l'exemple suivant :

$$\frac{d^2u}{dz^2} = \left(\frac{du}{dz}\right)^2 \frac{2u - 1}{u^2 + 1},$$

qui a la solution générale :

$$u = \tan [\ln (Az - B)] ;$$

A et B sont les constantes d'intégration, $z = B/A$ est la singularité essentielle mobile.

Le problème se pose alors de déterminer s'il existe ou non des équations différentielles du second ordre :

$$d^2u/dz^2 = F(z, u, du/dz),$$

où F est fonction analytique de z et rationnelle en u et du/dz telles que les points de branchement et singularités essentielles de toutes leurs solutions soient fixes.

Par une méthode de réductions successives on est conduit à cinquante types d'équations admissibles, tous étant intégrables au moyen de fonctions connues, sauf six ; les équations les plus intéressantes sont, bien entendu, celles, irréductibles, au nombre de six, qui servent à définir les transcendantes de Painlevé. On se bornera à écrire les deux plus simples :

$$d^2u/dz^2 = 6u^2 + z, \quad d^2u/dz^2 = 2u^3 + zu + a.$$

On démontre que les solutions de ces équations n'ont pas de singularité essentielle ni de points de branchement mobiles. Et mêmes elles n'ont pas de point de branchement et définissent ainsi des fonctions uniformes.

5. La théorie de la stabilité

Pour la description mathématique de très nombreux systèmes physiques oscillatoires on est conduit à des équations ou systèmes différentiels dont il convient de rechercher les solutions stationnaires ou périodiques et d'étudier leurs propriétés de stabilité.

Un modèle relativement simple est fourni par l'équation :

(45) $$dx/dt = Ax + f(x, t),$$

où $x \in \mathbf{R}^n$, A matrice $n \times n$ réelle et constante, $f(x, t)$ application continue de :

$$U \times [0, \infty]$$

dans $\mathbf{R}^n$, U étant un voisinage de l'origine, telle enfin que $f(0, t) = 0$. Il est clair que $x = 0$ est solution de (45) ou, comme l'on dit, un point critique. Mais que peut-on dire d'une solution dont la valeur initiale $x(0)$ est petite ? Sera-t-elle définie pour tout $t \geqslant 0$, et, dans l'affirmative, va-t-elle s'écarter notablement ou non de la solution d'équilibre $x = 0$. Cela amène à préciser le concept suivant de stabilité : une solution $x(t)$, du système $dx/dt = F(x, t)$ définie pour tout $t \geqslant t_0$ sera dite *stable* si, pour tout $\varepsilon > 0$, il existe $\delta = \delta(\varepsilon, t_0) > 0$ tel que toute autre solution $y(t)$ définie pour $t \geqslant t_0$ et vérifiant $\|y(t_0) - x(t_0)\| \leqslant \delta$ satisfait à $\|y(t) - x(t)\| \leqslant \varepsilon$ pour $t \geqslant t_0$.

Si, de plus, $y(t) - x(t) \to 0$ quand $t \to +\infty$, on dira que la solution $x(t)$ est *asymptotiquement stable.*

On observera que, pour discuter la stabilité d'une solution $x(t)$ d'un système quelconque $dx/dt = F(x, t)$, on pourra toujours, par le changement $x = x(t) + y$, se ramener à l'étude de la stabilité d'une solution stationnaire d'un système différentiel, ce qui justifie l'importance de système du type (45).

Revenant à ce cas, on peut énoncer le théorème (Poincaré-Liapounoff) : Si les valeurs propres de la matrice A ont toutes leur partie réelle négative et si la fonction $f(x, t)$ continue dans $\|x\| \leqslant \rho$, $t \geqslant 0$ est telle que :

$$\lim_{x \to 0} \sup_{t \geqslant 0} \|f(x, t)\| / \|x\| = 0,$$

$x = 0$ est solution asymptotiquement stable de (45).

Si l'une au moins des valeurs propres de A est à partie réelle positive, la solution $x = 0$ est instable.

Le théorème de stabilité demeure vrai si la matice A est une fonction continue périodique de t dont tous les coefficients caractéristiques sont à partie réelle négative.

Le problème fondamental de la stabilité de la solution $x = 0$ du système :

$$(46) \quad dx/dt = F(x, t), \quad \text{avec } F(0, t) = 0$$

peut être abordé par la méthode directe de Liapounoff.

Pour en expliquer le contenu, il faut donner quelques notations préliminaires : on dira que la fonction scalaire $V(x, t)$ a un signe constant si, dans un domaine $\Omega(a, \tau) : \|x\| \leqslant a, t \geqslant \tau$, convenable, elle est différentiable, ne prend que des valeurs d'un même signe ou nulle et $V(0, t) = 0$; elle sera dite positive ou négative selon la nature de ce signe.

Si $W(x)$ est une fonction scalaire indépendante du temps, on dira que $W(x)$ est définie positive (ou définie négative) si elle est différentiable, et est positive (ou négative) dans un $\Omega(a, \tau)$ convenable et ne s'annule qu'à l'origine. La fonction scalaire $V(x, t)$ sera dite définie positive (ou définie négative) s'il existe une fonction définie positive $W(x)$ telle que $V - W$ (ou $-V - W$) est positive dans un $\Omega(a, \tau)$ et $V(0, t) = 0$.

Supposant que $F(x, t)$ est continue dans un $\Omega(a, \tau)$ convenable, on a les théorèmes suivants :

– Si, pour le système (46) et dans un domaine $\Omega(a, \tau)$, il existe une fonction définie $V(x, t)$ dont la dérivée :

$$dV/dt = \partial V/\partial t + \sum_i F_i(x, t)\, \partial V/\partial x_i,$$

est d'un signe constant opposé, alors $x = 0$ est solution stable de (46).

– Si, pour le système (46) et dans un domaine $\Omega(a, \tau)$, $V(x, t)$ et dV/dt sont définies et de signe contraire et si

$$\lim_{x \to 0} \sup_{t \geqslant \tau} |V(x, t)| = 0,$$

alors $x = 0$ est une solution asymptotiquement stable de (46).

– Si, pour le système (46), on a pu construire une fonction $V(x, t)$ telle que :

$$\lim_{x \to 0} \sup_{t \geqslant \tau} |V(x, t)| = 0,$$

telle que dV/dt soit définie (positive ou négative) dans $\Omega(a, \tau)$, et que, pour chaque valeur $t > \tau$ et chaque $\eta > 0$ aussi petit qu'on veut, V et dV/dt puissent avoir, en certain point de $\|x\| < \eta$, le même signe, alors $x = 0$ est solution instable du système (46).

La difficulté d'application de ces théorèmes réside, bien entendu, dans la construction des fonctions de Liapounoff ; aussi de nombreux efforts ont été tentés visant, d'une part, à assouplir les condi-

tions imposées dans l'espoir de rendre plus facile cette construction, d'autre part, à reconnaître parmi ces conditions celles qui sont nécessaires pour tel type de stabilité.

Voici, pour conclure, un exemple d'application. Soit le système :

(47) $dx/dt = A(t)x + f(x, t)$,

où $A(t)$ est une matrice $n \times n$ fonction réelle continue et bornée dans $t \geqslant 0$, $f(x, t)$ application à valeurs dans $\mathbf{R}^n$, continue dans $\|x\| \leqslant a$, $t \geqslant 0$, et telle que

$$\lim_{x \to 0} \sup_{t \geqslant 0} \|f(x, t)\| \Big/ \|x\| = 0 ;$$

s'il existe une fonction de Liapounoff pour l'approximation linéaire $dx/dt = A(t)x$, c'est-à-dire une fonction $V(x, t)$ satisfaisant pour ce système linéaire aux exigences du troisième théorème énoncé plus haut, alors $x = 0$ est une solution asymptotiquement stable de (47).

6. Les solutions périodiques des systèmes différentiels

Un problème très important pour certaines applications est la recherche de solution périodique de système du type $dx/dt = f(x, t)$, où $f(x, t)$, application continue dans $\mathbf{R}^n$, est supposée périodique par rapport à la variable réelle t de période T (cas non autonome), ou encore du type $dx/dt = f(x)$ (cas autonome).

On ne dispose d'aucune méthode d'investigation assez puissante pour répondre à ces questions de manière générale. Les méthodes existantes sont de deux sortes. Les unes, méthode de perturbation, méthode de centrage, permettant l'étude de systèmes quasi linéaires, c'est-à-dire de systèmes dans lesquels la partie non linéaire apparaît multipliée par un paramètre qu'on suppose petit ; le calcul de représentations asymptotiques des solutions périodiques est généralement possible, ainsi que l'étude de la stabilité de ces solutions. Les autres sont des méthodes topologiques qui fournissent pour certains systèmes fortement non linéaires des résultats d'existence de solutions périodiques.

La méthode des perturbations (H. Poincaré)

Considérons l'équation :

(48) $x'' + x = \mu f(x, x', \omega t, \mu)$,

où x est une fonction scalaire, $x' = dx/dt$, $x'' = d^2x/dt^2$, f fonction périodique de t de période $2\pi/\omega$ et μ un petit paramètre, tous les éléments ainsi définis étant réels.

Quand $\mu = 0$ l'équation se réduit à $x'' + x = 0$ qui a pour solution générale $x = a \cos(t + \varphi)$, périodique de période 2π, a et φ désignant des constantes arbitraires.

Supposons que ω est voisin de l'unité ou mieux que $\omega^{-2} = 1 - \mu\eta$, η étant une fonction donnée de μ analytique dans le voisinage de 0,

$$\eta(\mu) = \sum_{k=0}^{\infty} \eta_k \mu^k.$$

Pour chercher si l'équation (48) a des solutions périodiques de période $2\pi/\omega$, nous posons $\omega t = \theta + \delta$, θ nouvelle variable, δ paramètre de translation. Il vient ainsi :

(49) $x'' + x = \mu g(x, x', \theta + \delta, \mu, \eta)$,
avec $x' = dx/d\theta$, $x'' = d^2x/d\theta^2$,

et :

$$g = \eta x + (1 - \mu\eta) f(x, x'/(1 - \mu\eta)^{1/2}, \theta + \delta, \mu),$$

et l'on recherche une solution périodique de (49) de période 2π en θ, telle que

$x(0) = a$, $x'(0) = 0$ (cette condition initiale fait comprendre le rôle du paramètre de translation δ). La théorie de Poincaré montre que si a et δ_0 sont solution réelle du système d'équations :

$$\int_0^{2\pi} \sin\tau \times f(a \cos\tau, -a \sin\tau, \tau + \delta_0, 0)\, d\tau = 0,$$

$$\pi\eta_0 a + \int_0^{2\pi} \cos\tau \times f(a \cos\tau, -a \sin\tau, \tau + \delta_0, 0)\, d\tau = 0,$$

telle que :

$$\int_0^{2\pi} \sin\tau \times (df/d\theta)(a \cos\tau, -a \sin\tau, \tau + \delta_0, 0)\, d\tau = 0,$$

et si $g(x, x', \theta + \delta, \mu, \eta)$ continue par rapport à tous ses arguments peut être développée en série de puissances de $x - a\cos\theta$, $x' + a\sin\theta$, μ, $\eta - \eta_0$, $\delta - \delta_0$ convergente pour tout θ si tous ces arguments sont de module moindre qu'un nombre positif ρ, alors il existe une solution périodique de l'équation (48) que l'on peut représenter par la série :

$$(50)\qquad x(\theta) = a\cos\theta + \sum_1^\infty \gamma_n(\theta)\,\mu^n.$$

Celle-ci peut être dérivée terme à terme deux fois, et :

$$(51)\qquad x'(\theta) = -a\sin\theta + \sum_1^\infty \gamma_n'(\theta)\,\mu^n,$$

$$x''(\theta) = -a\cos\theta + \sum_1^\infty \gamma_n''(\theta)\,\mu^n.$$

On peut représenter δ par une série :

$$(52)\qquad \delta = \delta_0 + \mu\delta_1 + \ldots$$

et l'on peut obtenir les développements (50) et (52) par substitution dans l'équation (49) et identification dans les deux membres des coefficients des mêmes puissances de μ. On obtient ainsi un système récurrent :

$$\gamma_n'' + \gamma_n = \varphi_n(\gamma_1, \ldots, \gamma_{n-1}, \delta_1, \ldots, \delta_{n-1}),\ n \geqslant 1.$$

On détermine les $\gamma_n(\theta)$ et les coefficients δ_k en imposant aux fonctions γ_n la condition de périodicité par rapport à θ de période 2π et en plus : $\gamma_n(0) = \gamma'_n(0) = 0$.

Cette méthode de calcul appliquée à l'équation de Duffing :

$$x'' + x + \mu x^3 = \mu h \cos\omega t,$$

avec $\omega^{-2} = 1 - \mu\eta$, h constante, conduit à :

$$\delta_0 = 0,\quad \eta_0 a + h - 3a^3/4 = 0.$$

On peut donc obtenir trois solutions périodiques, et en tout cas au moins une ; on a la représentation :

$$x = a\cos\theta + \mu(a^3/32)(-\cos\theta + \cos 3\theta) + O(\mu^2),$$
$$\theta = \omega t,\quad \omega^{-2} = 1 - \mu\eta_0 + O(\mu^2)$$

La théorie qui précède permet de donner une description satisfaisante du phénomène de synchronisation des oscillateurs quasi linéaires. Si l'oscillateur linéaire est attaqué par une force périodique non linéaire de période voisine de sa période propre, il y a synchronisation sur la force excitatrice, l'oscillation ayant en outre une amplitude bien définie.

La théorie de Poincaré permet aussi de rendre compte du phénomène de démultiplication de fréquence, c'est-à-dire l'existence de solutions sous-harmoniques, périodiques de période multiple de celle de la force excitatrice. Ces résultats d'une grande importance théorique et pratique sont spécifiques de la non-linéarité.

Le cas des systèmes autonomes peut être traité de façon similaire ; considérons par exemple l'équation :

$$(53) \qquad x'' + x = \mu f(x, x', \mu)$$

et proposons-nous de chercher s'il existe une solution périodique de (53) qui, lorsque $\mu \to 0$, tend vers une solution bien déterminée de $x'' + x = 0$, périodique de période 2π. Cela suggère que la période de la solution cherchée $2\pi/\omega$ peut être représentée par $\omega^{-2} = 1 - \mu\eta$, où $\eta(\mu)$ est une fonction de μ qui, à la différence du cas précédent, est une inconnue du problème. Avec cette représentation et la nouvelle variable $\omega t = \theta$, l'équation devient :

$$(54) \quad \begin{aligned} x'' + x &= \mu(\eta x + (1 - \mu\eta) f(x, x'/(1-\mu\eta)^{1/2}, \mu)) \\ &= \mu g(x, x', \mu, \eta), \end{aligned}$$

dont il convient de rechercher une solution périodique de période 2π. On peut alors démontrer suivant la méthode de Poincaré que si a_0 et η_0 sont des nombres réels satisfaisant aux équations :

$$\int_0^{2\pi} \sin\tau f(a_0\cos\tau, -a_0\sin\tau, 0)\, d\tau = 0,$$

$$\pi\eta_0 a_0 + \int_0^{2\pi} \cos\tau \times f(a_0\cos\tau, -a_0\sin\tau, 0)\, d\tau = 0,$$

tels que :

$$a_0 \int_0^{2\pi} \sin\tau \times [\cos\tau\, (\partial f/\partial x)(a_0\cos\tau, -a_0\sin\tau, 0) - \sin\tau\, (\partial f/\partial x')(a_0\cos\tau, -a_0\sin\tau, 0)]\, d\tau \neq 0.$$

Si, en outre, $g(x, x', \mu, \eta)$ peut être développé en série de puissances de $x - a_0 \cos\theta$, $x + a_0\sin\theta$, μ, $\eta - \eta_0$ convergente lorsque ces quantités sont de module assez petit, alors il existe une fonction $\eta(\mu)$ pour laquelle (54) possède une solution périodique $x(\theta)$ de période 2π telle que $x(0) = a_0$, $x'(0) = 0$, qu'on peut représenter par les séries :

$$\eta(\mu) = \eta_0 + \mu\eta_1 + \mu^2\eta_2 + \ldots$$
$$x(\theta) = a_0\cos\theta + \gamma_1(\theta)\mu + \ldots$$

Le développement de $x(\theta)$ peut être dérivé terme à terme deux fois et l'on peut obtenir les coefficients de ces séries par substitution dans (54) et identification dans les deux membres des termes de même puissance en μ.

La détermination de $\gamma_n(\theta)$ et η_k se fait alors par un procédé récurrent : $\gamma''_n + \gamma_n = \varphi(\gamma_1, \ldots, \gamma_{n-1}, \eta_1 \ldots, \eta_{n-1})$ et les conditions $\gamma_n(0) = 0$, $\gamma'_n(0) = 0$, avec $\gamma_n(\theta)$ périodique de période 2π.

Appliquant ces considérations à l'équation de Van der Pol :

$$x'' + \mu(x^2 - 1)x' + x = 0,$$

on trouve ainsi :

$$\frac{1}{\omega^2} = 1 + \frac{\mu^2}{8} + O(\mu^3), \quad \omega t = \theta,$$

$$x(\theta) = 2\cos\theta + \left(\frac{3}{4}\sin\theta - \frac{1}{4}\sin 3\theta\right)\mu + \left(-\frac{1}{8}\cos\theta + \frac{3}{16}\cos 3\theta - \frac{5}{96}\cos 5\theta\right)\mu^2 + O(\mu^3).$$

La méthode de centrage (Kryloff-Bogoliuboff-Haag)

a) Considérons le système :

$$(55) \qquad dx/dt = \mu f(x, t, \mu),$$

avec $x \in \mathbf{R}^n$, f application à valeurs dans $\mathbf{R}^n$, périodique par rapport à t de période T et pourvue de dérivées partielles premières continues. L'idée fondamentale est de substituer à (55) l'équation :

$$(56) \qquad dx/dt = \mu F(x)$$

avec : $\quad F(x) = (1/T)\int_0^T f(x, t, 0)\, dt.$

On peut établir que les solutions $x(t)$ et $\tilde{x}(t)$ respectivement de (55) et (56) qui prennent pour $t = 0$ même valeur demeurent assez voisines sur un intervalle de temps d'autant plus grand que μ est plus petit. De manière plus précise, sous certaines hypothèses, on établit qu'il existe une fonction croissante $h(t)$ telle que :

$$\|x(t) - \tilde{x}(t)\| \leqslant \mu h(\mu t).$$

Ce résultat signifie essentiellement que, choisissant μ assez petit, l'écart entre les solutions sera moindre qu'un nombre donné ε sur un intervalle de temps qui sera de l'ordre de μ^{-1}, donc très grand si μ est petit :

$$\mu < \varepsilon / h(\mathrm{L}) \Rightarrow \|x(t) - \tilde{x}(t)\| < \varepsilon,$$

pour $t \in [0, \mathrm{L}/\mu]$ et a fortiori pour $t \in [0, \mathrm{L}h(\mathrm{L})/\varepsilon]$,

b) D'un autre côté, cette méthode permet de prévoir aussi l'existence de solutions périodiques et d'étudier leur stabilité.

Plus précisément, on a le théorème : Si $a \in \mathbf{R}^n$ est tel que $\mathrm{F}(a) = 0$, et si la matrice :

$$(\partial \mathrm{F}_i / \partial x_j)(a) = (1/\mathrm{T}) \int_0^{\mathrm{T}} (\partial f_i / \partial x_j)(a, t, 0)\, dt$$

est non singulière, alors le système (55) possède une solution périodique de période T pour μ assez petit, qui, lorsque $\mu \to 0$, tend uniformément vers a ; de plus, si les valeurs propres de cette matrice ont toutes leurs parties réelles négatives, la solution périodique mise en évidence est asymptotiquement stable.

De nombreuses extensions sont possibles et pour conclure nous citerons un résultat fondamental : soit le système quasi linéaire :

$$(57) \quad dx/dt = \mathrm{A}x + f(t) + \mu \mathrm{F}(x, t, \mu),$$

où A est une matrice réelle $n \times n$ constante, $f(t)$ et $\mathrm{F}(x, t, \mu)$ des applications dans $\mathbf{R}^n$ périodiques en t de période T, F pourvue de dérivées partielles du premier ordre continues.

Si aucune valeur propre de la matrice A n'est égale à $\pm 2\pi iq/\mathrm{T}$, q entier, alors pour μ assez petit, le système (57) a une solution périodique unique de période T, asymptotiquement stable si les valeurs propres de A sont toutes à partie réelle négative.

Si certaines valeurs propres de A sont multiples de $\pm 2\pi i/\mathrm{T}$, alors le système $dx/dt = \mathrm{A}x$ possède un système maximal de solutions périodiques de période T, $\varphi^{(j)}(t)$ de composantes $\varphi_{ij}(t)$ avec $1 \leqslant j \leqslant m \leqslant n$. Le système adjoint : $dx/dt + {}^t\mathrm{A}x = 0$ (${}^t\mathrm{A}$ matrice transposée de A) possède aussi un système maximal de m solutions périodiques de période T, $\psi^{(k)}(t)$ de composantes $\psi_{ik}(t)$, qu'on peut toujours supposer construit de telle sorte que :

$$\langle \varphi^{(j)}, \psi^{(k)} \rangle = \sum_i \varphi_{ij}(t)\, \psi_{ik}(t) = \delta_{jk}.$$

Le système linéaire non homogène $dx/dt = \mathrm{A}x + f(t)$, sous réserve que l'on ait :

$$(58) \quad \int_0^{\mathrm{T}} \langle \psi^{(k)}(s), f(s) \rangle\, ds = \int_0^{\mathrm{T}} \Big(\sum_{j=1}^{n} \psi_{jk}(s) f_j(s) \Big) ds = 0,$$

pour $1 \leqslant k \leqslant m$, possède alors une famille de solutions périodiques de période T, qu'on peut écrire :

$$x(t) = \alpha_1 \varphi^{(1)}(t) + \ldots + \alpha_m \varphi^{(m)}(t) + \tilde{x}(t),$$

où les α_i sont des constantes arbitraires.

Revenant au système (57), on écrit les équations :

$$(59)\quad h_p(\alpha_1, ..., \alpha_m) = \int_0^T \left(\sum_{j=1}^{n} \psi_{jp}(s)\, F_j(x(s), s, 0) \right) ds = 0,$$

avec $1 \leqslant p \leqslant m$, on suppose qu'il existe un système de valeurs réelles $\alpha_1^0, ..., \alpha_m^0$ solution des équations (59) telles que le jacobien :

$$\frac{\partial(h_1, ..., h_m)}{\partial(\alpha_1, ..., \alpha_m)}$$

ne soit pas nul pour cet ensemble de valeurs.

Alors le théorème de Malkin affirme que, sous réserve des conditions (58), pour μ assez petit, le système (57) possède une solution périodique de période T qui, lorsque $\mu \to 0$ tend vers :

$$\bar{x}(t) = \alpha_1^0 \varphi^{(1)}(t) + ... + \alpha_m^0 \varphi^{(m)}(t) + \tilde{x}(t).$$

Il est généralement possible de déterminer les termes suivants du développement asymptotique de la solution par rapport à μ et d'en étudier la stabilité.

Ajoutons, pour conclure, que ces méthodes permettent aussi de discuter du cas des systèmes autonomes, et que d'importantes généralisations sont possibles, en particulier pour l'étude des systèmes presque périodiques.

Les méthodes topologiques

Considérons le système différentiel :

$$dx = f(x, t),$$

avec f à valeurs dans $\mathbf{R}^n$, périodique en t de période T. Sous certaines conditions, on peut définir la solution $x(t\,;x_0)$ de (59), qui, pour $t = 0$, prend la valeur x_0, valable pour $t \in [0, T]$.

Définissant l'opérateur $\mathfrak{I}x_0 = x(T\,;x_0)$ de $\mathbf{R}^n$ dans $\mathbf{R}^n$, l'on voit que, pour obtenir une solution périodique de période T de (59), le problème fondamental consiste à trouver un point fixe de cet opérateur, c'est-à-dire x_0 tel que : $x_0 = \mathfrak{I}x_0$.

On peut présenter cette idée de manière différente ; proposons-nous, par exemple, de rechercher la solution périodique de période T du système :

$$(60)\qquad dx/dt = A(t)x + g(x, t),$$

où $A(t)$, $g(x, t)$ sont respectivement matrice $n \times n$, applications dans $\mathbf{R}^n$ périodiques en t de période T. On suppose que le système homogène :

$$(61)\qquad dx/dt = A(t)x$$

n'a pas de solution périodique de période T.

On introduit, pour le système matriciel :

$$(62)\qquad dZ/dt = A(t)Z$$
$$(63)\qquad Z(0) = Z(T),$$

une matrice $n \times n$, $G(t, t')$ (matrice de Green) satisfaisant aux propriétés suivantes : $G(t, t')$ vérifie (62) pour $0 \leqslant t < t'$ et $t' < t \leqslant T$, satisfait à (63) relativement à t, enfin est discontinue en $t = t'$ avec un saut :

$$G(t' + 0, t') - G(t' - 0, t') = I.$$

Si $X(t)$ est la matrice résolvante de (62) alors $X(T) - I$ est inversible (car (61) n'a pas de solution périodique T) et l'on peut montrer que :

$$G(t, t') = \begin{cases} X(t)[I - X(T)]^{-1}X^{-1}(t'), & 0 \leqslant t' < t \leqslant T, \\ X(t)[I - X(T)]^{-1}X(T)X^{-1}(t'), & 0 \leqslant t < t' \leqslant T. \end{cases}$$

On s'assure facilement que toute solution périodique de période T de (60) est

solution de l'équation intégrale non linéaire :

$$(64) \quad x(t) = \int_0^T G(t,t')g(x(t'),t')\,dt',$$

et inversement.

Si l'on considère le second membre de (64) comme un opérateur $\mathfrak{J}$ agissant sur l'espace des fonctions continues $x(t)$, $t \in [0, T]$, on voit que le problème consiste à trouver un point fixe de cet opérateur, c'est-à-dire une fonction continue $x(t)$ telle que $x = \mathfrak{J}x$.

La théorie des points fixes des opérateurs non linéaires a reçu dans les dernières décennies un développement considérable et a rendu possible la recherche de solutions périodiques de systèmes différentiels dans de nombreux cas.

Nous allons, pour conclure, ajouter quelques développements dans le cas des systèmes autonomes de dimension 2.

Soit le système :

$$(65) \qquad \begin{aligned} dx/dt &= f(x,y) \\ dy/dt &= g(x,y), \end{aligned}$$

x, y, variables scalaires, f, g, fonctions réelles continues au sens de Lipschitz dans un ensemble D ouvert et borné du plan x, y.

Si C^+ (ou C^-) est une semi-orbite de (65), c'est-à-dire l'ensemble des points $P(t)$ dont les coordonnées sont $\varphi(t)$, $\psi(t)$, solution de (65) définie dans $t \geqslant t_0$ (ou $t \leqslant t_0$) pour un certain t_0, un point Q du plan sera dit point limite de C^+ (ou de C^-) s'il existe une suite de nombres réels t_n, $t_n \to +\infty$ (ou $t_n \to -\infty$) telle que $P(t_n) \to Q$, si $n \to \infty$. L'ensemble de tous les points limites d'une semi-orbite C^+ (ou C^-) est désigné par ($L(C^+)$ $L(C^-)$), et $L(C) = L(C^+) \cup L(C^-)$ est l'ensemble limite de l'orbite.

On a les théorèmes suivants :

– Si C^+ est une semi-orbite contenue dans un ensemble fermé $K \subset D$, alors $L(C^+)$ est non vide, fermé et connexe.

Définition. Les points de D en lesquels f et g s'annulent sont les points critiques ; tout autre point de D sera dit régulier.

– Soit C^+ une semi-orbite positive contenue dans un sous-ensemble fermé $K \subset D$ et supposons que $L(C^+)$ contienne un point régulier Q. Alors L'orbite C_Q passant par Q existe en tant qu'orbite complète $t \in]-\infty, +\infty[$ et $C_Q \subset L(C^+)$.

– Soit C^+ une semi-orbite positive contenue dans un sous-ensemble fermé $K \subset D$. Si $L(C^+)$ se compose exclusivement de points réguliers, alors C^+ ($= L(C^+)$) est une orbite périodique ou $L(C^+)$ est une orbite périodique ; dans le second cas, on dit que $L(C^+)$ est un cycle limite.

– *Corollaire* (*théorème de Poincaré-Bendixson*). Si C^+ est une semi-orbite contenue dans un sous-ensemble compact $K \subset D$, dans lequel il n'y a pas de point critique, alors K contient une orbite périodique.

Si T est la période et si :

$$h = \int_0^T \left(\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} \right) dt < 0,$$

cette intégrale étant calculée sur le cycle, celui-ci est asymptotiquement stable.

Voici, pour conclure, l'exemple suivant (équation de Liénard) : $x'' + x = f(x')$, qu'on écrira :

$$\begin{aligned} dx/dt &= y \\ dy/dt &= f(y) - x. \end{aligned}$$

On suppose $f(y)$ pourvue d'une dérivée continue et telle que $f(y) = g(y) - \alpha y$, α constante avec $dg/dy > 0$, $dg/dy(0) > \alpha$, $g(-y) = -g(-y)$, $|g(y)| < c$, c constant ; le théorème de Poincaré-Bendixson

peut être utilisé dans ce cas pour mettre en évidence un cycle limite.

Soit E un espace topologique ; un système dynamique est une famille d'opérateurs F_t ($t \in \mathbf{R}$) de E dans E possédant les propriétés suivantes :

a) F_0 est l'opérateur identité,

b) pour tous t_1, t_2 réels $F_{t_1+t_2} = F_{t_1}F_{t_2}$,

c) l'élément $F_t p$ de E dépend continûment de $(p, t) \in E \times \mathbf{R}$.

En particulier si l'on prend pour E l'espace $\mathbf{R}^n$ et si $f(x, t)$ est une fonction continue de $\mathbf{R}^n \times \mathbf{R}$ dans $\mathbf{R}^n$, lipschitzienne en x, on peut définir $x(t, x_0)$ la solution unique du système : $dx/dt = f(x, t)$, $x(0, x_0) = x_0$, puis la famille d'opérateurs $F_t : \mathbf{R}^n \rightarrow \mathbf{R}$ définie par $F_t(x_0) = x(t, x_0)$, laquelle constitue un système dynamique au sens de la définition donnée plus haut.

L'étude des systèmes dynamiques constitue donc un prolongement naturel de la théorie des équations différentielles et a donné lieu à de nombreux travaux : problèmes de stabilité, problèmes ergodiques (cf. théorie ERGODIQUE), etc.

Une autre généralisation importante consiste à considérer des équations différentielles dans lesquelles la fonction inconnue a des valeurs dans un espace métrique, par exemple un espace de Hilbert ou un espace de Banach.

On peut ainsi discuter des équations du type :

$$dx/dt = Ax, \quad dx/dt = Ax + f(x, t),$$

où x est pour toute valeur réelle de t élément d'un espace de Banach E, A désignant un opérateur linéaire de E dans E, $f(x, t)$ application non linéaire de $E \times \mathbf{R}$ dans E.

Dans le cas où l'opérateur A est indépendant de t, sous certaines conditions additionnelles, l'étude de l'équation $dx/dt = Ax$ conduit à la théorie bien développée des semi-groupes d'opérateurs dans E. Le cas des équations non linéaires a été aussi considéré en liaison avec les problèmes d'évolution, la question se posant, en général, de savoir comment se comportent pour $t \rightarrow +\infty$ les solutions de tels systèmes.

Le développement considérable de l'analyse fonctionnelle a beaucoup contribué aux progrès de ces théories dont l'étude se poursuit activement.

MAURICE ROSEAU

7. Intégration numérique des équations différentielles

Méthode d'Euler

Prenons d'abord le cas d'une équation différentielle du 1er ordre : Trouver y, fonction d'une variable x, dérivable sur $[x_0, x_0 + a] = I$, telle que, f désignant une fonction continue sur $I \times \mathbf{R}$,

$$y'(x) = f(x, y(x)),$$

pour tout $x \in I$, et :

$$y(x_0) = \lambda,$$

où λ est donné dans $\mathbf{R}$.

L'idée est de remplacer le problème théorique précédent, noté P, par le problème discrétisé P_n suivant (méthode d'Euler) : Trouver $Y_n = (y_0, y_1, ..., y_n)$, suite finie de $n + 1$ nombres réels telle que :

$$y_{i+1} = y_i + \frac{a}{b} f(x_i, y_i), \quad 0 \leqslant i \leqslant n - 1,$$

où :

$$x_i = x_0 + \frac{ia}{n}, \quad y_0 = \lambda ;$$

ce problème P_n est obtenu en divisant $I = [x_0, x_0 + a]$ en n parties égales avec un

pas égal à $h = a/n$ et en cherchant une approximation y_i de $y(x_i)$ où y est la solution (lorsqu'elle est unique) de P_1.

Le raisonnement, fort simple, est le suivant : Si y est solution unique de P_1, $y'(x_i)$ est proche de :

$$\frac{y(x_{i+1})-y(x_i)}{h}$$

et donc, si y_i est « proche » de $y(x_i)$, on peut remplacer P_1 par P_n.

P_n est un problème discrétisé associé à P_1. On remarque alors immédiatement qu'une notion importante va devoir être précisée : comment dire que la solution de P_n converge vers celle de P_1 lorsque n tend vers $+\infty$. L'analyse numérique devra fournir des majorations pour $|y(x_i)-y_i|$.

Dans la suite, nous supposerons toujours que f satisfait à la condition (L) suivante appelée *condition de Lipschitz globale* : il existe $L \geqslant 0$ tel que :

$$|f(x,u)-f(x,v)| \leqslant L|u-v|,$$

pour tout $x \in I$, $u \in \mathbf{R}$ et $v \in \mathbf{R}$.

Cette condition assure l'existence et l'unicité du problème P (cf. chap. 4).

On peut espérer que, si h est assez petit, le nombre y_i est proche de $y(x_i)$. C'est pourquoi le nombre $e_i = y(x_i) - y_i$ sera appelé l'erreur au point x_i, tandis que :

$$E_n = \max_{0 \leqslant i \leqslant n} |e_i|$$

est l'erreur globale aux points x_i, $0 \leqslant i \leqslant n$.

Le procédé P_h est dit *convergent* si :

$$\lim_{n \to +\infty} E_n = 0.$$

Notons que cette notion de convergence est peu habituelle ; car la solution de P est une fonction $x \mapsto y(x)$, alors que la solution Y_n de P_n est une suite finie de $n+1$ nombres réels. Ainsi, y et Y_n n'appartiennent pas au même espace et nous ne pouvons pas mesurer une distance éventuelle de y à Y_n. Néanmoins, cette notion de convergence satisfait le physicien, qui ne cherche pas explicitement y mais qui veut surtout des valeurs approchées de $y(x_i)$ pour un certain pas h. Si E_n est petit, l'erreur de discrétisation provoquée par le passage de P à P_n est faible.

Théorème. Si (L) est satisfaite, le procédé P_n est convergent. Plus précisément, on peut montrer que :

$$|e_i| \leqslant \frac{e^{L(x_i-x_0)}-1}{L} \max|y'(t_1)-y'(t_2)|$$
$$\leqslant \frac{e^{aL}-1}{L} \max|y'(t_1)-y'(t_2)|,$$

où les maximums sont pris pour $t_1 \in I$, $t_2 \in I$, avec $|t_1 - t_2| \leqslant h$.

Cette majoration prouve la convergence ; car, f étant continue, la solution unique de P possède une dérivée continue sur I, donc uniformément continue, de sorte que le maximum $y'(t_1) - y'(t_2)$, pour $|t_1 - t_2| \leqslant h$, tend vers 0 avec h.

Néanmoins, cette majoration ne donne pas de bornes pour l'erreur commise, car elle fait intervenir la dérivée de la fonction inconnue.

Cherchons une majoration de E_n. Prenons tout d'abord (L) comme seule hypothèse. On peut alors montrer que la solution unique de P satisfait à :

$$|y(x)-y_0| \leqslant \frac{K_1}{L}(e^{L(x-x_0)}-1),$$

où :

$$K_1 = \max_{x \in I}|f(x,y_0)|.$$

Une majoration de K_1 est en général assez simple à déterminer, de sorte qu'on peut définir un ensemble borné $\mathcal{D}$ de

$I \times \mathbf{R}$ qui contient $(x, y(x))$ tel que y soit solution de P. C'est l'ensemble $\mathcal{D}$, défini par :

$$|y(x) - y_0| \leqslant \frac{K_1}{L}(e^{L(x-x_0)} - 1), \quad x \in I.$$

Si, sur $\mathcal{D}$, qui est un ensemble borné fermé, la fonction f possède des dérivées partielles continues :

$$\frac{\partial f}{\partial x}, \ \frac{\partial f}{\partial y},$$

on peut alors majorer $|e_i|$ par :

$$\frac{e^{aL}-1}{L}\|y''\| \leqslant \frac{e^{aL}-1}{L} \times \left(\left\|\frac{\partial f}{\partial x}\right\|_{\mathcal{D}} + \|f\|_{\mathcal{D}} \cdot \left\|\frac{\partial f}{\partial y}\right\|_{\mathcal{D}}\right)h,$$

Remarquons que cette majoration est utilisable si on peut calculer aisément :

$$\frac{\partial f}{\partial x}, \ \frac{\partial f}{\partial y},$$

et si on peut majorer simplement :

$$\left\|\frac{\partial f}{\partial x}\right\|_{\mathcal{D}} = \max\left|\frac{\partial f}{\partial x}(x,y)\right|,$$

où le maximum est pris pour $(x, y) \in \mathcal{D}$, et de même pour :

$$\|f\|_{\mathcal{D}}, \ \left\|\frac{\partial f}{\partial y}\right\|_{\mathcal{D}}.$$

Le résultat final est alors :

$$E_n \leqslant Kh = K\frac{a}{n}.$$

Ainsi, la convergence est en $1/n$. On peut l'accélérer par la méthode d'extrapolation à la limite (méthode de Richardson, ou encore de Romberg ; cf. représentation et approximation des FONCTIONS).

La méthode d'extrapolation à la limite est très facile à appliquer. On peut, par exemple, diviser I en n parties égales, prendre pour $x = x_{i,0}$ un des points de subdivision, puis diviser en $2n$ parties égales en écrivant $x = x_{i(1)}$, puis en $4n$ partie égales en posant $x_{i(2)} = x$, puis en $2^p n$ parties égales en posant $x_{i(p)} = x$.

On résout alors les p schèmes :

$$y_{(k)+1} = y_{j(k)} + hf(x_{j(k)}, y_{j(k)}),$$

et on obtient $y_{i(k)}$ pour $1 \leqslant k \leqslant p$.

Si f est suffisamment différentiable sur $\mathcal{D}$, les conditions d'application de l'extrapolation à la limite sont satisfaites et on peut remplacer $y_{i(p)}$ par une moyenne convenable :

$$\tilde{y}_{i(p)} = \sum_{k=1}^{p} y_{i(k)} l_k,$$

telle que $\tilde{y}_{i(p)}$ converge vers $y(x)$ plus vite que $y_{i(p)}$ lorsque p tend vers $+\infty$.

Une famille de méthodes numériques

Nous avons jusqu'ici remplacé P par P_n :

$$y_{i+1} = y_i + hf(x_i, y_i), \quad y_0 = \lambda.$$

Nous proposons maintenant une généralisation de ce procédé : cette généralisation n'a pas pour unique objet de généraliser, elle prétend permettre d'aboutir à des méthodes plus efficaces que la méthode d'Euler.

Dans P intervient une fonction et f un nombre λ. Dans la méthode d'Euler, P_n faisait intervenir la même fonction f et le même nombre λ. Associons ici à f une fonction f_n et à λ un nombre λ_n, et remplaçons le problème P par le problème P_n. Pour le moment, f_n est une fonction donnée, définie sur $I \times \mathbf{R}$, et dépendant du paramètre h.

Considérons le problème P_h : Trouver $Y_n = (y_0, ..., y_n)$ tel qu'on ait :

$$y_{i+1} = y_i + hf_h(x_i, y_i), \quad y_0 = \lambda_h.$$

Faisons une étude de la convergence et essayons de choisir λ_h et f_h pour que l'« ordre de convergence » (cf. *infra*) soit bon.

Consistance

Nous dirons que le procédé P_h est *consistant* par rapport à P si :

$$\lim_{h \to 0} f_h(x,y) = f(x,y), \quad \forall x \in I, \forall y \in \mathbf{R}.$$

Cette condition est relativement simple à vérifier et assez intuitive : f_h doit « ressembler à f, pour h suffisamment petit ». Par exemple, dans la méthode d'Euler, on avait $f_h = f$, pour tout h, et cette méthode était donc trivialement consistante par rapport à P.

S'il en est ainsi, on peut démontrer que, pour toute solution y de $y' = f(x, y)$, on a, avec $h = a/n$,

$$\lim_{n \to \infty} \max_{0 \leqslant i \leqslant n-1} \left| \frac{1}{h} [y(x_{i+1}) - y(x_i)] - f_h(x_i, y(x_i)) \right| = 0.$$

Remarquons que cette notion ne fait intervenir que les f_h et non les λ_h.

Stabilité

Considérons les deux problèmes définis par la même équation $y' = f(x, y)$ et par les données initiales $y(x_0) = \lambda$ d'une part, et $y(x_0) = \tilde{\lambda}$ d'autre part. Supposons que (L) soit satisfaite. Les deux problèmes précédents possèdent respectivement une solution y et une solution $\tilde{y}$. Il peut arriver que, même lorsque $|\lambda - \tilde{\lambda}|$ est faible, la quantité :

$$\max_{x \in I} |y(x) - \tilde{y}(x)|$$

soit grande. C'est un phénomène d'instabilité du problème posé (indépendamment de toute méthode numérique de résolution). Ce phénomène d'instabilité est très gênant, car, si $\lambda = \pi$ par exemple, toute méthode numérique remplacera π par $\tilde{\lambda} = 3{,}141\,59...$ avec un nombre fini de décimales. Nous retrouverons évidemment cet ennui de résolution théorique dans la résolution numérique. Cela nous conduit à la définition suivante.

Un *procédé* P_h défini par f_h et λ_h est dit *stable* si, pour toute matrice triangulaire infinie (ε_{in}), $0 \leqslant i \leqslant n-1$, telle que :

$$\lim_{n \to +\infty} \max_{0 \leqslant i \leqslant n-1} |\varepsilon_{in}| = 0,$$

il existe des constantes K_1 et K_2 telles que :

$$\max_{0 \leqslant i \leqslant n} |y_i - z_i| \leqslant K_1 |y_0 - z_0| + K_2 \max_{0 \leqslant i \leqslant n-1} |\varepsilon_{in}|,$$

où y_i et z_i, $0 \leqslant i \leqslant n$, sont solutions de :

$$y_{i+1} = y_i + h f_h(x_i, y_i),$$

avec y_0 quelconque, et de :

$$z_{i+1} = z_i + h[f_h(x_i, y_i) + \varepsilon_{in}].$$

Cette condition de stabilité exprime le fait qu'une légère variation, sur les données initiales y_0 et z_0 d'une part, sur f_h d'autre part, n'entraîne pas de grosses variations sur les résultats obtenus en résolvant P_h. La condition de stabilité n'est pas simple. Voici une condition suffisante : le procédé P_h est stable s'il existe $\Lambda > 0$ tel que :

$$|f_h(x,u) - f_h(x,v)| \leqslant \Lambda |u - v|,$$

pour tous $x \in I$, $u \in \mathbf{R}$, $v \in \mathbf{R}$ et $h \in [0, h_0]$; donc Λ est indépendant de x, y, v et h. Nous admettrons ce résultat.

Convergence

Considérons les problèmes P et P_h suivants :

$$P : y' = f(x,y), y(x_0) = \lambda,$$
$$P_h : y_{i+1} = y_i + h f_h(x_i, y_i), y_0 = \lambda_h,$$

admettant pour solutions respectives les fonctions $y \in C^1(I)$ et $Y_n = (y_0, ... y_n)$.

Nous dirons que le schéma P_h est convergent si :

$$\lim_{n \to +\infty} E_n = 0,$$

où :

$$E_n = \max_{0 \leqslant i \leqslant n} |y(x_i) - y_i|.$$

Théorème fondamental de convergence. Si le schéma P_h est stable d'une part, et consistant par rapport à P d'autre part, le schéma P_h est convergent, à condition que :

$$\lim_{h \to 0} \lambda_h = \lambda.$$

En effet, comme P_h est consistant par rapport à P, on peut écrire :

$$\frac{1}{h}[y(x_{i+1}) - y(x_i)] - f_h(x_i, y(x_i)) = \varepsilon_{in},$$

où y est la solution de P et où :

$$\max_{0 \leqslant i \leqslant n-1} |\varepsilon_{in}|$$

tend vers zéro lorsque n tend vers $+\infty$.

En posant $z_i = y(x_i)$, on a :

$$z_{i+1} = z_i + h[f_h(x_i, z_i) + \varepsilon_{in}],$$

de sorte que, par la condition de stabilité, on obtient :

$$E_n = \max_{0 \leqslant i \leqslant n} |y_i - y(x_i)|$$
$$\leqslant K_1 |\lambda - \lambda_h| + M_2 \max_{0 \leqslant i \leqslant n-1} |\varepsilon_{in}|.$$

Si λ_h tend vers λ, on a :

$$\lim_{n \to +\infty} E_n = 0,$$

Ordre d'une méthode

Nous venons de voir que, si :

$$\lim_{h \to 0} f_h(x,y) = f(x,y),$$
$$|f_h(x,u) - f_h(x,v)| \leqslant \Lambda |u - v|,$$

alors, le procédé P_h défini par f_h est stable et consistant, donc convergent : cela prouve simplement que :

$$\lim_{n \to 0} E_n = 0,$$

Nous dirons qu'une méthode est d'ordre p si on a $E_n \leqslant Kh^p$ lorsque $\lambda_h = \lambda$. Des conditions pour qu'une méthode soit d'ordre p ont été données : elles font intervenir les dérivées par rapport à h de la fonction $h \mapsto f_h(x, y) = g(h)$. Les techniques utilisées sont relativement simples : il suffit de faire un usage fréquent de la formule de Taylor. Contentons-nous de donner certains résultats.

Supposons que f est p fois continûment différentiable. Posons :

$$\mathcal{F}_0 = f, \quad \mathcal{F}_1 = \frac{\partial \mathcal{F}_0}{\partial x} + f \frac{\partial \mathcal{F}_0}{\partial y}$$

et, de proche en proche,

$$\mathcal{F}_{k+1} = \frac{\partial \mathcal{F}_k}{\partial x} + f \frac{\partial \mathcal{F}_k}{\partial y}, \quad 0 \leqslant k \leqslant p.$$

Il en résulte que, si y est solution de P, on a :

$$y^{(k+1)}(x) = \frac{d^k}{dx^k} f(x, y(x))$$
$$= \mathcal{F}_k(x, y(x)), \quad 0 \leqslant k \leqslant p.$$

Pour que le procédé P_h soit d'ordre p, il suffit que :

$$\lim_{h \to 0} \frac{d^k g}{dh^k} = \frac{1}{k+1} \mathcal{F}_k, \quad 0 \leqslant k \leqslant p - 1.$$

La *méthode du développement de Taylor* consiste à prendre :

$$f_h(x,y) = f(x,y) + \frac{h}{2} \mathcal{F}_1(x,y)$$
$$+ \dots + \frac{h^{p-1}}{p!} \mathcal{F}_{p-1}(x,y);$$

on peut alors démontrer que $E_n \leqslant Kh^p$. Cette méthode est cependant assez difficile à utiliser car elle nécessite le calcul effectif des $\mathcal{F}_k$.

On peut aussi écrire :

$$f_h(x,y) = a_1 f(x,y) + a_2 f(x + a_3 h, y + a_4 h f(x,y))$$

et on cherche a_1, a_2, a_3, a_4 pour que la méthode soit d'ordre aussi élevé que possible. En écrivant que :

$$f_0(x,y) = f(x,y), \quad \frac{dg}{dh} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right),$$

on obtient les conditions $a_1 = 1 - \alpha$, $a_2 = \alpha$, $a_3 = a_4 = 1/(2\,\alpha)$.

Pour $\alpha = 1/2$, on a la *méthode de Heun* :

$$f_h(x,y) = \frac{1}{2}[f(x,y) + f(x + h, y + hf(x,y))].$$

Pour $\alpha = 1$, on a la *méthode d'Euler améliorée* :

$$f_h(x,y) = f\left(x + \frac{h}{2}, \quad y + \frac{h}{2}f(x,y)\right).$$

Indiquons enfin la *méthode de Runge et Kutta* ; on pose :

$$f_h(x,y) = \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4],$$

avec :

$$k_1 = f(x,y),$$
$$k_2 = f(x + \frac{h}{2}, y + \frac{h}{2}k_1),$$
$$k_3 = f(x + \frac{h}{2}, y + \frac{h}{2}k_2),$$
$$k_4 = f(x + h, y + hk_3);$$

cette méthode est d'ordre 4.

Méthode des approximations successives

On sait que le problème P est équivalent au problème Q : Trouver y continu sur I tel que :

$$y(x) = \lambda + \int_{x_0}^{x} f(t, y(t))\,dt.$$

Si y est solution de Q, et donc de P, on a :

$$y(x_{p+k}) - y(x_p) = \int_{x_p}^{x_{p+k}} f(t, y(t))\,dt.$$

Supposons connues des valeurs approchées $y_{p-q}, \dots, y_p$ de $y(x_{p-q}), \dots, y(x_p)$. Nous connaissons alors des valeurs approchées de : $f(x_{p-j}, y(x_{p-j}))$, $0 \leqslant j \leqslant q$. Posons :

$$f_{p-j} = f(x_{p-j}, y_{p-j}), \quad 0 \leqslant j \leqslant q,$$

et considérons le polynôme d'interpolation sur les points $x_{p-q}, \dots, x_p$ de la fonction $t \mapsto f(t, y(t))$ dont nous connaissons des valeurs approchées f_{p-j} aux points x_{p-j}. Le calcul approché de :

$$\int_{x_p}^{x_{p+k}} f(t, y(t))\,dt$$

peut se faire par une formule d'intégration de type ouvert, n'utilisant que les points précédant x_p et le point x_p.

On a, par exemple, la formule approchée :

$$y(x_{p+1}) \simeq y(x_p) + \int_{x_p}^{x_{p+1}} \mathrm{P}(t)\,dt,$$

où P désigne le polynôme d'interpolation sur les points $x_{p-q}, \dots, x_p$.

Considérons donc le schéma ou algorithme défini par la méthode d'intégration de Newton-Gregory :

$$y_{p+1} = y_p + h\sum_{j=0}^{q} c_j \nabla^j f_p, \quad q \leqslant p \leqslant n - 1,$$

où les *différences régressives* sont définies à partir de :

$$\nabla f_1 = f_1 - f_0, \dots, \quad \nabla f_k = f_k - f_{k-1};$$

par récurrence :

$$\nabla^1 = \nabla, \quad \nabla^j f_p = \nabla^{j-1} f_p - \nabla^{j-1} f_{p-1}.$$

Ce schéma nous permet de calculer y_{p+1}, pour $p = q$, puisque les $\nabla^j f_q$ sont connus ; y_{q+1} étant connu, nous calculons y_{q+2} en changeant p en $p + 1$ et ainsi de

suite, en utilisant toujours le même nombre $q+1$ de points qui précèdent x_{p+1}. Le problème qui se pose est alors de savoir si les y_p ainsi calculés sont proches des valeurs $y(x_p)$ de la solution exacte.

L'algorithme précédent est dit explicite ; car, à chaque étape, il définit une inconnue y_{p+1} explicitement. La méthode suppose connues les valeurs de départ $y_0, ..., y_q$ qui sont des valeurs approchées de $y(x_0), ..., y(x_q)$ et qui ont donc été obtenues par une autre méthode, par exemple une méthode à un pas, alors qu'ici nous obtenons y_{q+1} par une méthode utilisant $y_0, ..., y_q$, et c'est pourquoi nous dirons que la méthode précédente est une méthode multipas, à q pas exactement.

Remarquons que les c_j peuvent être déterminés par la méthode de la fonction génératrice introduite à propos de l'intégration numérique par la méthode de Newton-Gregory.

On peut aussi utiliser la méthode suivante. On calcule :

$$\int_{x_p}^{x_{p+1}} f(t, y(t))\, dt$$

de façon approchée par une formule de type fermé qui utilise donc y_{p+1} que l'on cherche. Le schéma s'écrit :

$$y_{p+1} = y_p + h \sum_{j=0}^{q} d_j \nabla_j f_{p+1},$$

où les d_j peuvent aussi être calculés par la méthode de la série génératrice.

Cette équation s'écrit encore sous la forme :

$$y_{p+1} = hf(x_{p+1}, y_{p+1}) + \mathrm{K}_p,$$

où K_p ne dépend que des point x_{p-k}, pour $k \geqslant 0$.

On obtient donc y_{p+1} implicitement ; mais, p étant fixé, la détermination de y_{p+1} peut être faite par itération : on prend la valeur y_{p+1}^0 obtenue par le schéma explicite et on itère en écrivant :

$$y_{p+1}^{(m)} = hf(x_{p+1}, y_{p+1}^{(m-1)}) + \mathrm{K}_p.$$

La suite des $y_{p+1}^{(m)}$, pour $m = 0, 1, 2, 3, ...$, converge vers la solution y_{p+1} du schéma implicite lorsque h est suffisamment petit ; car, si on écrit le schéma implicite sous la forme $y_{p+1} = \Phi(y_{p+1})$, on a :

$$\Phi(u) - \Phi(v) = h[f(x_{p+1}, u) - f(x_{p+1}, v)].$$

Donc on a la majoration :

$$|\Phi(u) - \Phi(v)| \leqslant h\mathrm{L}|u - v|$$

et la convergence est assurée si $h\mathrm{L} < 1$.

La première méthode explicite permet de définir une première approximation de y_{q+1} : nous dirons que c'est une *formule de prédiction*. La deuxième méthode permet de modifier le résultat précédent et de l'améliorer : c'est une *formule de correction*.

Ne cherchons pas à justifier a priori une méthode en utilisant un polynôme d'interpolation ou un autre et écrivons brutalement un schéma multi-pas sous la forme générale suivante.

On suppose que $y_0, y_1, ..., y_{q-1}$ sont obtenus par une méthode à un pas convergente ; on écrit :

$$(\mathrm{S}_p) \quad \sum_{i=0}^{q} \alpha_i y_{p-i} = h \sum_{i=0}^{q} \beta_i f_{p-i}, \quad q \leqslant p \leqslant n,$$

où les α_i et β_i sont des constantes données, indépendantes de h, de p et de f.

Dans la suite, nous chercherons des valeurs des α_i et des β_i pour que la méthode soit aussi efficace que possible. Nous supposerons que $\alpha_0 \neq 0$.

Pour $p = q$, on obtient :

$$(\mathrm{S}_q) \quad \sum_{i=0}^{q} \alpha_i y_{q-i} = h \sum_{i=0}^{q} \beta_i f_{q-i}.$$

Si $\beta_0 = 0$, le système (S_q) définit explicitement y_q par une formule du type :

$$y_q = \sum_{i=1}^{q} \alpha'_i y_{q-i} + h \sum_{i=1}^{q} \beta'_i f_{q-i},$$

puisque $y_0, ..., y_{q-1}$ sont connus ; de proche en proche, on obtient $y_q, y_{q+1}, ..., y_n$ en calculant à chaque étape y_q par une formule du type :

$$y_q = h \frac{\beta_0}{\alpha_0} f_q + K_{q-1},$$

où K_{q-1} est connu. Si h est suffisamment petit, on démontre, comme précédemment, que l'équation précédente (en y_q) possède une solution unique qui peut être obtenue par itération ; il suffit en effet que :

$$h L \left| \frac{\beta_0}{\alpha_0} \right| < 1 ;$$

de proche en proche, on obtient y_p, pour $q \leqslant p \leqslant n$, en résolvant chaque S_p.

On peut encore définir les notions de consistance (par rapport au problème posé) et de stabilité. Contentons-nous de donner certains résultats de convergence.

Nous dirons que la méthode précédente est *convergente* si :

$$E_n = \max_{0 \leqslant i \leqslant n} |y(x_i) - y_i|$$

tend vers 0 lorsque N augmente indéfiniment.

Cela suppose donc que les y_i, pour $0 \leqslant i \leqslant q-1$, ont été définis par une méthode à un pas convergente.

Considérons alors les polynômes :

$$A(z) = \sum_{i=0}^{q} \alpha_i z^{q-i},$$

$$B(z) = \sum_{i=0}^{q} \beta_i z^{q-i}.$$

Ces polynômes A et B définissent la méthode multi-pas, puisque S_p est défini par les α_i et par les β_i.

Une *condition suffisante* de convergence est donnée par le théorème suivant.

Théorème. Si $A(1) = 0$ et si $A'(1) = B(1)$, si de plus les racines z_i de A satisfont à $|z_i| \leqslant 1$, pour $1 \leqslant i \leqslant q$, et si les racines de module égal à 1 sont toutes simples, alors la méthode définie par les polynômes A et B est convergente.

On peut obtenir aussi des conditions pour que la méthode soit d'ordre k.

CHRISTIAN COATMELEC et *E.U.*

Bibliographie

V. I. ARNOLD, *Équations différentielles ordinaires*, M.I.R., Moscou, 1978 ; *Chapitres supplémentaires de la théorie des équations différentielles ordinaires*, *ibid.*, 1980 / M. BRAUN, *Differential Equations and their Applications : An Introduction to Applied Mathematics*, Springer-Verlag, New York-Berlin, 3e éd. 1986 / M. CROUZEIX & A. .L. MIGNOT, *Analyse numérique des équations différentielles*, 2e éd. rev., Masson, 1989 / J.-P. DEMAILLY, *Analyse numérique et équations différentielles*, Presses univ. Grenoble, 1991 / J. K. HALE, *Ordinary Differential Equations*, Krieger, Melbourne (Fla.), 2e éd. 1980 / H. REINHARD, *Équations différentielles : fondements et applications*, Dunod, 2e éd. 1989.

DIOPHANTIENNES

APPROXIMATIONS

La théorie des approximations diophantiennes concerne principalement l'approximation des irrationnels par des rationnels. Dans le cas d'un seul irrationnel, un rôle essentiel est joué par les fractions continuées (utilisées dès 1650 par Huygens pour le calcul des engrenages des horloges astronomiques).

L'approximation des irrationnels algébriques fut étudiée par une méthode directe en 1844 par Liouville ; ses résultats furent améliorés à de nombreuses reprises jusqu'à l'important et définitif résultat de Roth en 1955.

Dans le cas de plusieurs irrationnels, on peut soit chercher à approcher chacun d'eux par un rationnel, soit chercher à rendre minimale une forme linéaire à variables entières, à coefficients irrationnels (problème dual du précédent). Dans les deux cas, la théorie des réseaux de points (ou **Z**-modules, comme $\mathbf{Z}^n$ par exemple) joue un grand rôle, avec la caractérisation de ses bases et le théorème fondamental de Minkowski sur les domaines convexes symétriques d'un réseau ; ce dernier théorème conduit principalement à la résolution en entiers d'inégalités à coefficients irrationnels, ce qui est aussi un problème d'approximation diophantienne.

L'étude de la répartition modulo 1 a été également rattachée à cet article, étant encore, dans une certaine mesure, une question d'approximation diophantienne.



1. Z-modules et réseaux

Un **Z**-*module de* $\mathbf{R}^n$ est un ensemble $\mathcal{M}$ de points M de $\mathbf{R}^n$, de coordonnées $(x^1, x^2, ..., x^n)$, qui est sous-groupe additif de $\mathbf{R}^n$ (donc, s'il contient M′ et M″, il contient $u\,\mathrm{M}' + v\,\mathrm{M}''$ pour tout u et v de **Z**). On appelle *base* de $\mathcal{M}$ un ensemble $\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_r$ d'éléments de $\mathcal{M}$, tel que tout élément de $\mathcal{M}$ s'écrit, d'une manière unique, sous la forme $a^1\mathrm{A}_1 + a^2\mathrm{A}_2 + ... + a^r\mathrm{A}_r$, où $a^i \in \mathbf{Z}$. On remarquera qu'on peut avoir $r > n$ (par exemple dans **R** avec les nombres de la forme $a + b\sqrt{2}$, où a et b sont entiers, on a $r = 2$ pour $n = 1$).

Lorsque chaque point de $\mathcal{M}$ est *isolé* dans $\mathbf{R}^n$, c'est-à-dire est centre d'une boule ne contenant pas d'autre point de $\mathcal{M}$, le **Z**-module est dit *discret* ; cette propriété est évidemment caractérisée par le fait qu'il existe une boule de $\mathbf{R}^n$, de centre O, qui ne contient que O comme point de $\mathcal{M}$. On appelle *réseau* tout **Z**-module discret et on démontre qu'un réseau de $\mathbf{R}^n$ ne peut avoir de base comprenant plus de n éléments. Plus précisément, si un réseau admet une base de r éléments, l'espace vectoriel qu'il engendre est de dimension r.

Pour $n = 1$, tout réseau est donné par $x = mx_0$ où $m \in \mathbf{Z}$ et $x_0 = \inf|x|$ pour $x \in \mathcal{M} - \{\,\mathrm{O}\,\}$; en effet, $\mathcal{M}$ étant discret, x_0 existe bien, est non nul et appartient à $\mathcal{M}$, et tout x de $\mathcal{M}$ en est multiple, sans quoi le reste du quotient de $|x|$ par x_0 contredirait l'hypothèse faite sur x_0. Pour n quelconque, la démonstration de $r \leqslant n$ se fait par récurrence en projetant sur $\mathbf{R}^{n-1}$ parallèlement à l'un des vecteurs OA_i.

On supposera, sans rien restreindre dans ce qui suit, que les réseaux envisagés correspondent à $r = n$. Un exemple fondamental est celui des points à coordonnées entières de $\mathbf{R}^n$, auquel on peut toujours se ramener dès qu'on a une base $\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_n$ du réseau : c'est $\mathbf{Z}^n$.

Il est important de pouvoir caractériser les bases de $\mathbf{Z}^n$; on démontre qu'une condition nécessaire et suffisante pour que n points $\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_n$ de $\mathbf{Z}^n$ forment une base de ce réseau est que le déterminant de leurs coordonnées soit égal à ± 1. Dans le cas de $n = 2$, si (p_1, q_1) et (p_2, q_2) sont les coordonnées de A_1 et A_2, on doit donc avoir, pour une base, $p_1q_2 - p_2q_1 = \pm 1$, c'est-à-dire que les deux fractions p_1/q_1 et p_2/q_2 sont *adjacentes*.

On remarque enfin que OA_1A_2 forment base de $\mathbf{Z}^2$ si, et seulement si, le parallélogramme construit sur $\overrightarrow{OA_1}$, $\overrightarrow{OA_2}$ ne contient aucun point du réseau en son intérieur ; il a alors pour surface 1, et cela se généralise à $\mathbf{Z}^n$.

Un important théorème de Minkowski, sur les réseaux, sera vu plus loin.

2. Approximations d'un irrationnel. Fractions continuées

Dans le plan affine d'axes Ox, Oy, de vecteurs de base $\overrightarrow{OA}$, $\overrightarrow{OB}$, soit la demi-droite (OD) d'équation $x = \tau y$, avec $y \geqslant 0$ et $\tau \in \mathbf{R}$. Approcher τ par des rationnels p/q (avec $q > 0$) revient à approcher (OD) par des points du réseau de base $\overrightarrow{OA}$, $\overrightarrow{OB}$. Un point P(p, q) de ce réseau est un *point de voisinage à droite* pour (OD) si $p/q > \tau$ et $0 < p'/q' - \tau < p/q - \tau$ entraîne $q' > q$. Même définition *à gauche* avec $\tau > p/q$ et $0 < \tau - p'/q' < \tau - p/q$. Un point P($p$, q) est un *point réduit*, relativement à (OD), si $|p' - \tau q'| < |p - \tau q|$ entraîne $q' > q$.

Si τ est rationnel, soit $\tau = u/v$, la demi-droite (OD) porte le point entier P(u, v) et il n'y a plus, au-delà de P, ni de point de voisinage, ni de point réduit pour (OD). Les points antérieurs à P sont donnés par le théorème suivant, qui s'applique sans limitation lorsque τ est irrationnel : Si $P_{n,k}$ et $P_{n,k+1}$ sont deux points de voisinage consécutifs (pour q croissant), d'un même côté de (OD), alors :

a) La demi-droite portant le vecteur $\overrightarrow{P_{n,k}P_{n,k+1}}$ rencontre (OD) en un point D_{n+2} (non entier).

b) $\overrightarrow{P_{n,k}P_{n,k+h}} = h\,\overrightarrow{P_{n,k}P_{n,k+1}}$ donne, pour $h = 1, 2, ...,$ les points de voisinage suivant $P_{n,k}$, du même côté de (OD), jusqu'au dernier avant D_{n+2}, soit P_{n+2}. Ces points sont adjacents deux à deux.

c) $\overrightarrow{OP_{n+1}} = \overrightarrow{P_{n,k}P_{n,k+1}}$ et $\overrightarrow{P_{n+2}P_{n+1,1}} = \overrightarrow{P_{n,k}P_{n,k+1}}$ donnent deux points de voisinages consécutifs, P_{n+1} et $P_{n+1,1}$, de l'autre côté de (OD), ce qui permet de poursuivre l'opération et d'obtenir tous les points de voisinage au-delà de $P_{n,k}$, sur deux lignes polygonales appelées *lignes polygonales de Klein* relatives à (OD) ; ces lignes forment enveloppe convexe des points entiers situés de part et d'autre de (OD).

d) Seuls les sommets P_n de ces lignes polygonales sont des points réduits.

Ce théorème s'établit par des considérations géométriques très élémentaires. On pose en général $\overrightarrow{P_{n-2}D_n} = \alpha_n\,\overrightarrow{OP_{n-1}}$ et $\overrightarrow{P_{n-2}P_n} = a_n\,\overrightarrow{OP_{n-1}}$; donc $a_n = [\alpha_n]$, partie entière de α_n. La similitude des triangles OP_nD_n et $D_{n+1}P_{n-1}O$ donne alors $\alpha_{n+1} = 1/(\alpha_n - a_n)$ et l'on obtient ainsi, géométriquement, le *développement en fraction continuée* (régulière) de τ :

$$\tau = a_0 + \frac{1}{\alpha_1}, \quad \alpha_1 = a_1 + \frac{1}{\alpha_2}, ..., \quad \alpha_n = a_n + \frac{1}{\alpha_{n+1}}, ...,$$

avec $a_0 = [\tau]$ et $a_n = [\alpha_n]$. Les α_n sont les *quotients complets* du développement et les a_n, les *quotients incomplets*.

On écrit alors $\tau = [a_0, a_1, ..., a_{n-1}, \alpha_n]$ et $p_n/q_n = [a_0, a_1, ..., a_{n+1}, a_n]$, n-nième *réduite* (qui correspond au point P_n), d'où :

$$\begin{cases} p_n = p_{n-2} + a_n p_{n-1} \\ q_n = q_{n-2} + a_n q_{n-1}, \end{cases}$$

ce qui exprime que $\overrightarrow{P_{n-2}P_n} = a_n\,\overrightarrow{OP_{n-1}}$. C'est grâce à ces formules de récurrence qu'on calcule les p_n et les q_n connaissant les a_k.

Les résultats essentiels de la théorie de ces fractions continuées sont les suivants :
(1) On a :

$$\tau - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n(q_{n-1} + \alpha_{n+1} q_n)},$$

ce qui montre que l'approximation de τ par une réduite est d'autant meilleure que α_{n+1} est grand.

Par exemple $\pi = [3, 7, 15, 1, 292, 1, 1, 1, ...]$ donne une excellente approximation classique $p_3/q_3 = 355/113$.
(2) Une condition nécessaire et suffisante pour que deux irrationnels τ et α présentent, à partir de certains indices, les mêmes développements, est qu'ils soient liés par une transformation *homographique modulaire*, c'est-à-dire $\sigma = (a\tau + b)/(c\tau + d)$ avec $ad - bc = \pm 1$ et a, b, c et d entiers.
(3) Une condition nécessaire et suffisante pour que τ présente un développement périodique est que τ soit un irrationnel algébrique du second degré (théorème dû à Lagrange).

(4) $$|\tau - p_n/q_n| < 1/q_n^2,$$

pour tout n.

Sur deux réduites consécutives, l'une au moins vérifie :

$$|\tau - p_n/q_n| < 1/(2 q_n^2).$$

Sur trois réduites consécutives, l'une au moins vérifie :

$$|\tau - p_n/q_n| \leqslant 1/(\sqrt{5}\, q_n^2).$$

(5) Si $|\tau - u/v| < 1/(2 v^2)$,

u/v est une réduite du développement de τ.
(6) Si on développe le rationnel a/b en fraction continuée et si $a/b = p_n/q_n$, la réduite précédente $(p_{n-1})/(q_{n-1})$ fournit une solution de l'équation de Bezout :

$$ax - by = \pm 1.$$

(7) Si l'irrationnel $\sqrt{d}$ (d entier $\geqslant 2$) est développé en fraction continue, on a :

$$\sqrt{d} = [a_0, a_1, a_2, ..., a_{n-1}, 2a_0],$$

périodique à partir de a_1, période de n termes, avec $a_1 = a_{n-1}$, $a_2 = a_{n-2}$, ... (cela caractérise les développements de $\sqrt{d}$).

Si n est *pair*, les solutions de l'équation de Pell $x^2 - dy^2 = 1$ sont données par $x = p_{kn-1}$, $y = q_{kn-1}$, pour $k = 1, 2, 3, ...$ et $x^2 - dy^2 = -1$ n'a pas de solution.

Si n est *impair* les formules précédentes donnent : pour $k = 1, 3, 5, 7, ...$, les solutions de $x^2 - dy^2 = -1$, et, pour $k = 2, 4, 6, ...$, les solutions de $x^2 - dy^2 = 1$. Par exemple $x^2 - 13y^2 = 1$ conduit à :

$$\sqrt{13} = [3, 1, 1, 1, 1, 6],$$

d'où 649/180 pour $k = 2$, donnant la plus petite solution $x = 649$, $y = 180$ qu'il était difficile de trouver par essais successifs.

Ces résultats s'établissent à partir de la formule :

$$\tau = \frac{p_{n-2} + \alpha_n p_{n-1}}{q_{n-1} + \alpha_n q_{n-1}},$$

relation modulaire liant τ et α_n et exprimant que D_n est sur (OD). On en tire en effet aussitôt la formule de (1) et le résultat de (2), qui, géométriquement, signifie qu'à partir d'une autre base du réseau $\mathbf{Z}^n$ les points de voisinage et les points réduits sont, assez loin, les mêmes qu'à partir de la base OA, OB ; cela en raison du caractère géométrique intrinsèque des lignes polygonales de Klein. La condition nécessaire de (3) est évidente ($\alpha_k = \alpha_{k+\tau}$ donne α_k lié homographiquement – et même modulairement – à lui-même, donc racine d'une équation du second degré ; il en est de même pour τ). Sachant que τ est irrationnel du second degré, on peut

démontrer la condition suffisante. Algébriquement, en utilisant les équations du second degré que vérifient les α_n, équations à discriminant constant (ce qui permet, compte tenu de leurs signes, d'en limiter les coefficients ; d'où répétition et période). On peut aussi le démontrer géométriquement, à partir de la *rotation hyperbolique*. Cette transformation consiste à associer à (OD) la demi-droite (OD′) correspondant au conjugué τ' de τ et à remarquer que l'affinité d'axes (OD) et (OD′), et de multiplicateurs respectifs ν' et ν, où ν est une unité de $\mathbf{Q}[\tau]$ de la forme $u + c_2 v \tau$ (u et v entiers, c_2 coefficient du premier terme de l'équation définissant τ), conserve le réseau $\mathbf{Z}^2$ et, par conséquent, fait « glisser » sur elles-mêmes, à partir d'un point suffisamment éloigné, les lignes polygonales de Klein relatives à (OD). Il y a donc période (et période dès le début si $\tau > 1$ et $-1 < \tau' < 0$).

Les propriétés de (4) s'établissent en posant :

$$\frac{1}{q_n q_{n+1}} = \lambda_n \left[\left(\frac{1}{q_n} \right)^2 + \left(\frac{1}{q_{n+1}} \right)^2 \right]$$

et en montrant que λ_n ou λ_{n-1} est inférieur à $1/\sqrt{5}$, ce qui donne le résultat, grâce à :

$$\left| \frac{p_n}{q_n} - \tau \right| + \left| \tau - \frac{p_{n+1}}{q_{n+1}} \right| = \frac{1}{q_n q_{n+1}}.$$

La propriété (5) est évidente géométriquement, en considérant les parallélogrammes centrés en O, de côtés parallèles aux axes et dont un sommet est un point réduit.

La propriété (6) n'est autre que la récurrence fondamentale :

$$p_{n-1} q_n - p_n q_{n-1} = (-1)^n.$$

La propriété (7) est plus délicate à établir par le calcul ; l'étude de

$$\alpha_n = (b_n + \sqrt{\mathrm{D}})/c_n$$

conduit en effet à :

$$\alpha_{km-1} = [\sqrt{\mathrm{D}}] + \sqrt{\mathrm{D}}$$

et permet de conclure $p^2 - q^2 \mathrm{D} = (-1)^{km}$ pour la réduite d'ordre $(km - 1)$. Reste à montrer qu'on a là toutes les solutions.

Signalons d'autre part que la propriété (2), qui n'est plus vraie lorsque la transformation n'est pas modulaire, est remplacée, lorsque $ad - bc = m \geqslant 2$ par des formules assez simples de transformation des quotients incomplets, à condition que ceux-ci soient périodiques modulo m. C'est ainsi que le développement de

$$\frac{e+1}{e-1} = [2, 6, 10, ..., 2(2n+1), ...]$$

se transforme en :

$$e = [2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, ..., 2n, 1, 1, ...].$$

Pour obtenir le développement de $(e+1)/(e-1)$, Gauss a utilisé le développement en fraction continuée non régulière des séries hypergéométriques.

3. Approximations des irrationnels algébriques

On dit qu'un irrationnel τ est rationnellement approchable à l'ordre α s'il existe une constante dépendant de τ, soit $\mathrm{K}(\tau)$, telle que :

$$|\tau - p/q| < \mathrm{K}(\tau)/q^{\alpha}$$

ait une infinité de solutions.

On voit sans peine qu'un rationnel u/v est approchable à l'ordre 1 et pas au-delà. D'autre part, les propriétés des fractions

continuées montrent que tout irrationnel est approchable à l'ordre 2 au moins et qu'un irrationnel quadratique est approchable à l'ordre 2 et pas au-delà (à cause de la périodicité du développement). Ce dernier résultat est un cas particulier du théorème de Liouville (1844) relatif aux irrationnels algébriques de degré n : si τ est de degré n, il n'est pas approchable à un ordre supérieur strictement à n. En effet, si $f(\tau) = 0$, où f est le polynôme de degré n définissant τ, l'étude de :

$$f\left(\frac{p}{q}\right) - f(\tau) = \left(\frac{p}{q} - \tau\right) f'(\lambda)$$

donne élémentairement :

$$\left|\frac{p}{q} - \tau\right| > \frac{1}{Aq^n},$$

pour tout rationnel p/q.

Le théorème de Liouville a une grande importance historique, puisqu'il a permis de définir explicitement les premiers nombres *transcendants* (nombres de Liouville), grâce à des développements (décimaux ou en fraction continuée) lacunaires tels que :

$$x = 10^{-1!} + 10^{-2!} + ... + 10^{-n!} + ...$$
$$\text{ou : } x = [0, 10^{1!}, 10^{2!} ..., 10^{n!}, ...] ;$$

jusque-là on ne connaissait que l'existence des nombres transcendants (par complémentarité dans **R** des nombres algébriques) et ce n'est qu'en 1873 que Hermite établit la transcendance de e, permettant à Lindemann d'établir celle de π en 1882.

Le résultat de Liouville a été successivement amélioré par Thue (1908), établissant $\alpha \leqslant (n/2) + 1$, par Siegel (1921) $\alpha \leqslant 2\sqrt{n}$, par Dyson (1947) $\alpha \leqslant \sqrt{2}\, n$, et, en 1955, à l'aide d'une démonstration très technique, Roth améliorait définitivement le théorème : Tout irrationnel algébrique τ est approchable à l'ordre 2 et pas au-delà. Cela permet d'affirmer par exemple que le développement décimal lacunaire $x = 10^{-1} + 10^{-3} + 10^{-9} + ... + 10^{-3^n} + ...$ représente un nombre transcendant.

La recherche, pour un nombre algébrique τ, de la plus petite constante $k(\tau)$, pour laquelle :

$$\left|\tau - \frac{p}{q}\right| < \frac{k(\tau)}{q^2}$$

a une infinité de solutions, est alors intéressante. Le *nombre d'or* :

$$\alpha = \frac{1}{2}(\sqrt{5} - 1) = [1, 1, 1, ...]$$

a pour réduites les fractions de Fibonacci :

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{3}, \frac{3}{5}, \frac{5}{8}, \frac{8}{13}, \frac{13}{21}, ...$$

et on voit aisément que :

$$\left|\alpha - \frac{p_n}{q_n}\right| \sim \frac{1}{\sqrt{5}}\left(\frac{1}{q_n}\right)^2,$$

ce qui montre que $k(\alpha) = 1/\sqrt{5}$ (résultat de Hurwitz).

Mais, si l'on excepte le nombre α et ceux qui lui sont équivalents par transformation modulaire, on peut établir que :

$$k(\tau) \leqslant 1/(2\sqrt{2}).$$

Cela rejoint d'ailleurs les chaînes de Markoff-Hurwitz, qui, pour les irrationnels τ, étudient la limite supérieure $M(\tau)$ des constantes c telles que $|\tau - (p/q)| = 1/(cq^2)$; on obtient $M(\tau) = \sqrt{5}$ pour le nombre d'or et ses équivalents, puis $M(\tau) \geqslant 2\sqrt{2}$ pour les autres irrationnels, avec $M(\tau) = 2\sqrt{2}$ pour τ équivalent à $1 + \sqrt{2}$, puis $M(\tau) \geqslant \sqrt{221}/5$ pour les autres irrationnels, et ainsi de suite, les valeurs successives de $M(\tau)$ apparaissant étant données par la formule :

$$M(\tau) = \sqrt{9 - 4/u^2},$$

où $u = 1, 2, 5, 13, 29, \ldots$ est choisi de telle sorte que $u^2 + v^2 + w^2 = 3\,uvw$ soit résoluble en entiers.

Citons encore deux résultats sur les approximations asymétriques. Ségré établit, en 1946, que, pour tout $r \geqslant 0$ et pour tout irrationnel τ, il existe une infinité d'approximations p/q telles que :

$$\frac{-1}{\sqrt{1+4r}}\,\frac{1}{q^2} < \frac{p}{q} - \tau < \frac{r}{\sqrt{1+4r}}\,\frac{1}{q^2};$$

le cas $r = 0$ donne la propriété (4) des fractions continuées et le cas $r = 1$ donne le résultat de Hurwitz. Robinson établit en 1947 que, pour tout $\varepsilon > 0$ et pour tout irrationnel τ, il existe une infinité d'approximations p/q telles que :

$$\frac{-1}{\sqrt{5}-\varepsilon)}\,\frac{1}{q^2} < \frac{p}{q} - \tau < \frac{r}{\sqrt{5}+1)}\,\frac{1}{q^2};$$

cela montre qu'on peut renforcer une des inégalités du théorème d'Hurwitz sans affaiblir beaucoup l'autre.

4. Approximations simultanées

Étant donné k irrationnels $\tau_1, \tau_2, \ldots, \tau_k$, on peut soit chercher à les approcher par des fractions p_1/r, p_2/r, ..., p_k/r de même dénominateur (pas obligatoirement toutes irréductibles), soit chercher à rendre

$$u_1\tau_1 + u_2\tau_2 + \ldots + u_k\tau_k - w$$

minimum pour des entiers u_i et w. Ces deux problèmes duals l'un de l'autre sont également délicats. Le premier problème a été étudié initialement par Hermite, le second par Dirichlet. Une variante non homogène du deuxième problème consiste à rendre

$$u_1\tau_1 + u_2\tau_2 + \ldots + u_k\tau_k - \sigma - w$$

minimum, σ étant donné non entier.

Un algorithme de Jacobi généralise pour les irrationnels l'algorithme des fractions continuées. Il correspond, pour $k = 2$, à :

$$\begin{cases} \alpha_n = a_n + \dfrac{\beta_{n+1}}{\alpha_{n+1}} \\ \beta_n = b_n + \dfrac{1}{\alpha_{n+1}} \end{cases}$$

où $a_n = [\alpha_n]$ et $b_n = [\beta_n]$, avec $a_0 = [\tau]$ et $b_0 = [\sigma]$.

Cela, géométriquement, ramène le premier problème d'approximation simultanée à l'exploration des points de $\mathbf{Z}^3$ autour de la demi-droite (OD) portant le vecteur de composantes $(\tau, \sigma, 1)$. On obtient une suite de points liés par la récurrence $\overrightarrow{OP_n} = \overrightarrow{OP_{n-3}} + b_n\,\overrightarrow{OP_{n-2}} + a_n\,\overrightarrow{OP_{n-1}}$ et beaucoup de formules généralisent ce qui a été vu pour les fractions continuées. Malheureusement, si la convergence des réduites peut s'établir d'une manière générale, $|p - \tau r|$ et $|q - \sigma r|$ ne tendent pas toujours vers zéro. D'autre part, si on veut essayer de définir des points de voisinage, on doit tenir compte simultanément de $|\tau - p/r|$ et $|\sigma - q/r|$, ce qui peut se faire par leur borne supérieure, leur somme, la somme de leurs carrés, et bien d'autres manières qui ne sont pas équivalentes entre elles. C'est pourquoi Hermite a pu dire que ce problème n'avait cessé, durant cinquante ans, de le préoccuper et de le désespérer.

Lorsqu'un développement de Jacobi pour $k = 2$ est périodique, il est facile de voir que τ et σ sont deux éléments, non liés linéairement, d'un même corps cubique. La réciproque n'a pu être établie jusqu'à présent. Tout au plus peut-on dire, avec David (1956), que pour certains algorithmes voisins (où l'on prend par exemple $a'_n = a_{n+1}$ ou $b'_n = b_n + 1$) cette réciproque est inexacte. Il n'en reste pas moins que d'intéressants résultats ont été obtenus par Perron grâce à l'algorithme de Jacobi,

sur l'approximation de n entiers algébriques d'un corps de degré $(n+1)$.

Indépendamment de tout algorithme et par une simple application du « principe des tiroirs » de Dirichlet (Si $n+1$ objets sont dans n tiroirs, l'un au moins de ces tiroirs contient plus d'un objet) on démontre qu'il y a au moins une solution au système $|\tau_i - p_i/r| < 1/r^{1+\varepsilon}$, où $\varepsilon = 1/k$. Ce résultat de Kronecker est sans grand intérêt dès que k dépasse 3. Par dualité, on en déduit que :

$$|u_1\tau_1 + u_2\tau_2 + \ldots + u_k\tau_k - y| \leqslant 1/t^k$$

a des solutions entières u_i et y, non toutes nulles, avec $\sup |u_i| = t$.

Une étude plus précise de Khintchine lie l'indice ω_1 de $u_1\tau_1 + u_2\tau_2 + \ldots + u_k\tau_k - y$ (c'est-à-dire la borne supérieure des ω tels que cette forme soit approchable à $1/t^{k+\omega}$ près) à l'indice ω_2 des k nombres τ_i (c'est-à-dire la borne supérieure des ω tels que chaque τ_i soit approchable à $1/t^{1+(1+\omega)/k}$ près).

On a :

$$\omega_1 \geqslant \omega_2 \geqslant \frac{\omega_1}{k^2 + \omega_1(k-1)}.$$

Le cas non homogène (étude de $|\tau p - q - \sigma|$, ou, plus généralement, de $|\tau_1 u_1 + \tau_2 u_2 + \ldots + \tau_k u_k - w - \sigma|$) a permis à Tchebycheff, Kintchine, Kronecker, Hermite et Minkowski d'obtenir des résultats analogues à ceux du cas homogène.

Signalons enfin d'intéressantes études sur l'approximation d'un nombre complexe $\alpha + i\beta$ par le quotient P/Q de deux entiers de Gauss (Hermite, en liaison avec les formes quadratiques, Minkowski, Perron, Hurwitz en particulier). Un résultat essentiel est qu'il y a une infinité de solutions à

$$|\alpha + i\beta - (P/Q)| < 1/(\sqrt{3}|Q|^2).$$

Le théorème de Thue-Siegel-Dyson-Roth sur l'approximation d'un irrationnel algébrique a été généralisé au cas de plusieurs irrationnels algébriques, en 1970, par W. Schmidt. Ainsi, pour $\tau_1, \ldots, \tau_n$ des nombres algébriques tels qu'aucune combinaison linéaire $a_1\tau_1 + \ldots + a_n\tau_n$, avec $a_1, \ldots, a_n$ rationnels non tous nuls, ne soit un nombre rationnel, et pour ε réel positif arbitraire, il n'y a qu'un nombre fini d'entiers $p_1, \ldots, p_n, q$ $(q > 0)$ satisfaisant :

$$|\tau_i - p_i/q| < 1/q^{1+1/n+\varepsilon}.$$

Plus généralement, désignant par $\|x\|$ la distance d'un réel x au plus proche entier, pour $\tau_1, \ldots, \tau_n$ et ε comme ci-dessus, il n'y a qu'un nombre fini d'entiers positifs q tels que :

$$q^{1+\varepsilon}\|q\tau_1\| \ldots \|q\tau_n\| < 1,$$

et il n'y a qu'un nombre fini d'entiers $q_1, \ldots, q_n$ non nuls tels que :

$$|q_1 \ldots q_n|^{1+\varepsilon}\|q_1\tau_1 + \ldots + q_n\tau_n\| < 1.$$

5. Théorème de Minkowski et applications

Dans sa *Géométrie des nombres*, Minkowski établit en 1910 l'important théorème : Soit dans $\mathbf{R}^n$ un domaine S convexe, borné, symétrique par rapport à O et de volume supérieur à 2^n (ou égal à 2^n si ce domaine est fermé). Ce domaine contient au moins un point entier distinct de O (il contient donc aussi son symétrique par rapport à O).

La démonstration utilise l'homothétique S′ de S dans l'homothétie (0, 1/2). Pour un entier m assez grand, soit le réseau $(\mathbf{Z}/m)^n$ des points de coordonnées $x_i = u_i/m$ où $u_i \in \mathbf{Z}$. Soit N(m) le nombre d'hypercubes de côtés $1/m$ de ce réseau qui sont dans S′. On a N(m) $\times$ m^s aussi

voisin qu'on veut de $1/2^n$. V(S) pour m assez grand et cela permet d'affirmer l'existence de deux sommets où les u_i' et u_i'' sont congrus modulo m, pour $i = 1, 2, ..., n$. Le vecteur joignant ces deux sommets est donc un point entier, car $(u_i' - u_i'')/m$ est entier, qui appartient à $S' + S' = S$.

L'application essentielle de ce théorème concerne la résolution des systèmes d'inéquations diophantiennes :

$$|\alpha_i^1 x_1 + \alpha_i^2 x_2 + ... + \alpha_i^n x_n| \leqslant \lambda_i,$$

pour $i = 1, 2, ..., n$, où les α_i^j sont réels donnés, ainsi que les λ_i, et dont on cherche des solutions entières non banales (x_i entiers non tous nuls). Ces inéquations définissent une jauge de Minkowski (c'est ainsi qu'on appelle les domaines S définis ci-dessus) dont le volume est $2^n \lambda_1 \lambda_2 ... \lambda_n / |\Delta|$, où Δ est le déterminant des α_i^j. On peut donc affirmer que, si $\lambda_1 \lambda_2 ... \lambda_n \geqslant |\Delta|$, il y a des solutions entières, autres que le point O, au système des n inéquations. En particulier, on retrouve le résultat de Kronecker : le système de n inéquations

$$|\tau_i - p_i/r| \leqslant 1/(r^{1+\varepsilon})$$

est résoluble pour $\varepsilon = 1/n$; en effet, cela s'écrit :

$$\begin{cases} |r\tau_i - p_i| \leqslant (1/r^{\varepsilon}) \\ |r| \leqslant r, \end{cases}$$

système de $(n + 1)$ inéquations résoluble puisque la condition de Minkowski est ici réalisée :

$$(1/r^{\varepsilon})^n \times r = 1, \quad \text{si} \quad \varepsilon = 1/n.$$

On peut d'ailleurs, si les α_i^j sont entiers, appliquer ces résultats à des systèmes d'équations linéaires, car :

$$u_i^1 x_1 + ... + u_i^n x_n = 0$$

équivaut à :

$$|u_i^1 x_1 + ... + u_i^n x_n| \leqslant \lambda_i < 1,$$

si les u_i^j sont entiers et si l'on cherche à résoudre en entiers x_i.

Le théorème de Minkowski s'étend de plus au cas complexe, à condition que les inéquations du système soient réelles ou imaginaires conjuguées 2 à 2 (avec alors le même λ_i pour ces deux inéquations). Il vient encore la condition $\lambda_1 \lambda_2 ... \lambda_n \geqslant |\Delta|$, suffisante pour entraîner l'existence de solutions entières autres que le point O. C'est sous cette forme que le théorème permet la démonstration d'un important théorème de Dirichlet sur l'existence des unités dans une extension algébrique de **Q**.

On peut aussi appliquer le théorème des jauges de Minkowski en définissant celles-ci par :

$$|L_1| + |L_2| + ... + |L_n| < \lambda$$

où $L_i = \alpha_i^1 x_1 + ... + \alpha_i^n x_n$. On démontre ainsi que les deux inéquations :

$$|L_1| + |L_2| + ... + |L_n| \leqslant (n!|\Delta|)^{1/n}$$

et

$$|L_1 L_2 ... L_n| \leqslant n^{-n} n! |\Delta|$$

ont l'une et l'autre des solutions non banales, car :

$$n |L_1 L_2 ... L_n|^{1/n} \leqslant |L_1| + |L_2| + ... + |L_n|.$$

On peut toutefois noter que, contrairement au cas des inéquations, où la condition $|\lambda_1 \lambda_2 ... \lambda_n| \geqslant |\Delta|$ ne peut être améliorée d'une manière générale, on peut parfois améliorer les résultats ci-dessus. Par exemple, pour $n = 2$, on peut affirmer que l'inéquation $|L_1 L_2| \leqslant |\Delta| \sqrt{5}$ admet toujours des solutions non banales.

La réduction des formes quadratiques utilise aussi le théorème de Minkowski, appliqué à $L_1^2 + L_2^2 + ... + L_n^2$.

Blichfeldt (1914) a étendu à d'autres domaines que des jauges les méthodes de Minkowski ; ces recherches ont été poursuivies par Mordell, Davenport et Mahler.

6. Répartition modulo 1

Quoiqu'il ne s'agisse pas à proprement parler d'approximation diophantienne, on peut ranger dans cet article l'étude des suites de nombres réels, modulo 1. Il s'agit, pour une suite (u_n), de la répartition sur $[0, 1[$ de $\{u_n\} = u_n - [u_n]$ où $[u_n]$ est la partie entière de u_n.

Ce n'est qu'en 1884 que Kronecker établit que, si θ est irrationnel, ses multiples $n\theta$ sont, modulo 1, partout denses sur $[0, 1[$. Cela signifie que, quel que soit $x \in [0, 1[$ et quel que soit $\varepsilon > 0$, il existe une infinité de valeurs de n pour lesquelles $|\{n\theta\} - x| < \varepsilon$. En effet, $\{n_1\theta\}$ est différent de $\{n_2\theta\}$ si $n_1 \neq n_2$; il existe donc au moins un point d'accumulation des nombres $(n\theta)$, c'est-à-dire qu'on peut trouver n_1 et n_2 avec $(n_1 - n_2)\,\theta \in]0, \varepsilon[$, d'où les multiples $m(n_1 - n_2)\theta$ qui fournissent des points, modulo 1, à moins de ε de tout x de $[0, 1[$.

On remarquera que le problème de la répartition sur un cercle des points d'abscisse curviligne $n\theta$ conduit au même résultat si θ est incommensurable à π (ici on raisonne modulo 2π). De même, par exemple, l'étude des premiers chiffres du nombre 2^n, écrit en base 10, conduit à étudier la mantisse de $n \log 2$, c'est-à-dire sa répartition modulo 1. Comme log 2 est irrationnel, puisque $10^{p/q} \neq 2$, on en déduit qu'on peut toujours trouver une infinité de valeurs de n telles que 2^n commence par k chiffres quelconques imposés.

La notion d'*équirépartition* fut mise au point par Weyl en 1916. La suite (u_n) est dite équirépartie modulo 1 si les $\{u_n\}$ sont denses sur $[0, 1]$ et si, de plus, pour tout $[\alpha, \beta] \subset [0, 1]$ le nombre $\varphi_N(\alpha, \beta)$ d'indices n pour lesquels $n \leqslant N$ et $\{u_n\} \in [\alpha, \beta]$ vérifie :

$$\varphi_N/N \to \beta - \alpha \quad \text{quand } N \to +\infty ;$$

une condition suffisante pour que la suite $(f(n))$ soit dense sur $[0, 1[$ est que :

$$f(x) \to +\infty \quad \text{et} \quad f(x+1) - f(x) \to 0 \quad \text{quand } x \to +\infty ;$$

c'est ainsi que la suite $(\theta \operatorname{Log}^a n)$ est dense sur $[0, 1[$ modulo 1, quel que soit le nombre réel θ non nul et le nombre réel $a > 1$.

Dès 1912, Bohl, Sierpinski et Weyl établissent l'équirépartition de $(n\theta)$ pour θ irrationnel cependant que Fejer donne des conditions suffisantes d'équirépartition ou de non-équirépartition : Si f est strictement croissante, à dérivée continue monotone, avec $f(x) \to +\infty$, $f'(x) \to 0$, $xf'(x) \to \infty$ quand $x \to +\infty$, il y a équirépartition. Si au contraire $xf'(x) \to 0$, il n'y a pas équirépartition. On en déduit les résultats concernant $(\theta \operatorname{Log}^a n)$ (équirépartition si $a > 1$, non-équirépartition si $a \leqslant 1$). En 1916, Weyl énonce le critère d'équirépartition : celle-ci est caractérisée, pour une suite $f(n)$, par le fait que, pour tout entier h non nul,

$$\sum_{k=1}^{n} \exp 2i\pi h f(k)$$

est un $o(n)$ pour $n \to \infty$.

Plus tard (1933), Koksma établit que la suite λt^n, où λ est réel non nul fixé, est équirépartie modulo 1 pour presque tous les $t > 1$ (mais on ne connaît aucun t pour lequel on ait établi cette équirépartition ;

on pense par exemple que $(3/2)^n$ est équiréparti modulo 1, mais on n'a pas pu le démontrer jusqu'ici). En revanche, une catégorie importante de nombres algébriques échappe à cette équirépartition : il s'agit des nombres de Pisot-Vijayaragavan, qui sont des entiers algébriques θ tels que $\theta > 1$, les conjugués θ_i, pour $i = 2, 3, ..., s$ (s est le degré de θ), étant tous en modules inférieurs à 1. Il s'ensuit que θ^n converge vers zéro modulo 1 (raisonner sur $\theta^n + \theta^n_2 + ... + \theta^n_s$ qui est un entier). Salem a démontré en 1944 que l'ensemble S des nombres de Pisot était fermé.

MARCEL DAVID

Bibliographie

A. BAKER, *Transcendental Number Theory*, Cambridge Univ. Press, 1990 / G. H. HARDY & E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Oxford Univ. Press, New York, 5e éd. 1979 / S. LANG, *Introduction to Diophantine Approximations*, Addison, Reading (Mass.), 1966 / W. M. SCHMIDT, *Diophantine Approximations and Diophantine Equations*, Springer-Verlag, New York, 1991 / K. B. STOLARSKY, *Algebric Numbers and Diophantine Approximation*, M. Dekker, New York, 1974 / G. WUSTHOLZ, *Diophantine Approximation and Transcendence Theory*, in *Lecture Notes in Mathematics Ser.*, vol. 1290, Springer-Verlag, New York.

DIOPHANTIENNES ÉQUATIONS

Diophante d'Alexandrie, vers les années 250 de notre ère, fut le premier à rechercher systématiquement les solutions en nombres entiers, ou rationnels, d'une équation ou d'un système d'équations polynomiales à coefficients entiers. Bien que ce ne soit qu'avec Fermat (1601-1665) que les méthodes utilisées pour résoudre ces équations prirent un aspect vraiment arithmétique, c'est-à-dire faisant pleinement intervenir la factorisation des nombres entiers, une longue tradition appelle équation diophantienne la donnée d'un système d'équations polynomiales à coefficients entiers :

$$f_1(x_1, ..., x_n) = 0$$
$$\cdots\cdots$$
$$f_r(x_1, ..., x_n) = 0,$$

à résoudre en nombre entiers, ou rationnels, $x_1, ..., x_n$.

Selon que l'on veut résoudre en nombre entiers ou rationnels, les méthodes et les résultats diffèrent souvent sensiblement.

Des méthodes générales existent pour résoudre un système d'équations du premier degré, ou encore *une* équation du second degré. On dispose encore de méthodes pour étudier *une* équation du troisième degré, mais déjà, là, les problèmes ouverts abondent. Quant aux équations de degré supérieur, il est significatif que beaucoup d'ouvrages consacrés aux équations diophantiennes n'apparaissent que comme une accumulation de résultats disparates.

De fait, il a maintenant été établi (J. Robinson, Yu. V. Matijasevic, 1970) que le dixième problème de Hilbert a une réponse négative : il n'existe pas d'algorithme universel permettant de décider si une équation diophantienne a une solution en nombre *entiers*.

On ne peut donc espérer obtenir des méthodes générales que pour des types particuliers de systèmes d'équations. Comment classifier ces « types » ? La façon la plus évidente est d'utiliser le degré des équations définissant le système. Cette classification est souvent trop grossière, mais peut être affinée grâce à la géométrie algébrique. Cette dernière permet d'obte-

nir des résultats généraux parfois difficiles à traduire en termes d'équations concrètes. La géométrie algébrique nous donne aussi la mesure de notre ignorance : ainsi aucun changement de variables ne permet de ramener une équation du type :

$$ax^5 + by^5 + cz^5 + d = 0$$

(a, b, c, d entiers non nuis). à résoudre en (x, y, z) nombres rationnels, à un type d'équations que l'on sait actuellement traiter.

Par extension. on appelle aussi équations diophantiennes des équations dans lesquelles les exposants figurent parmi les inconnues ; la plus fameuse équation de ce type est :

$$x^m - y^n = 1,$$

à résoudre en entiers (x, y, m, n) au moins égaux à 1, qui n'admettrait (E. Catalan, 18 14-l 894) que la solution :

$$3^2 - 2^3 = 1.$$

De grands progrès ont été réalisés dans cette direction.

Dans cet article, l'ensemble des entiers naturels est désigné par N, l'anneau des entiers relatifs par **Z.** le corps des nombres rationnels par Q.



1. Le premier et le second degré

Le premier degré

L'équation :

$$ax + by = c,$$

ou a, b, c sont entiers relatifs, se traite classiquement,

Si c n'est pas divisible par le plus grand commun diviseur de a et b, il n'y a pas de solution entière : on peut donc supposer a et b premiers entre eux et utiliser la résolution de $au + bv = 1$ (Bezout), d'où $x = u_0c + kb$. $y = v_0c \quad ka$, avec u_0 et v_0 solution particulière de l'équation de Bezout et k entier relatif quelconque.

La solution (u_0, v_0) peut se trouver par essais successifs, si a et b ne sont pas trop grands ; sinon, on développe a/b en fraction continuée et. si $a/b = p_n/q_n$ est la n-ième réduite, on prend la (n 1)-ième qui, au signe près, donne $u_0 = q_n$, et $v_0 = -p_{n\ 1}$. Par exemple, si $355\ x + 113\ y = 1$. on a $355/113 = [3, 7, 16]$, d'où $p_{n-1}/q_n, = 22/7$ et $u_0 = \quad 7$. $v_0 = 22$. Cela correspond aussi, si l'on veut, à l'application de l'algorithme d'Euclide au couple (a, b).

L'équation :

$$a_1x_1 + a_2x_2 + ... + a_nx_n = c,$$

que nous écrirons A X $= c$ avec :

$$A = (a_1, a_2\ , ..., \ a\ ,)\quad \text{e t}\quad X\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

se résout, en supposant les a_i premiers entre eux dans leur ensemble, par les points d'un réseau à (II 1) dimensions (cf. approximations DIOPHANTIENNES. chap. 1) :

$$X = c\,U_0 + \lambda_1 B_1 + \lambda_2 B_2 + ... + \lambda_{n-1} B_{n-1},$$

où U_0 est solution particulière de l'équation de Bezout A X = 1 et où B_1, B_2, B_{n-1} engendrent le module des solutions de A . X = 0.

Un *système non homogène* :

$$A_iX = c_i$$

(i = 1, 2, ..., r) se discutera dans $\mathbf{Z}^r$, où on l'écrira :

$$\sum_1^n x_i V_i = c,$$

avec V, et C vecteurs colonnes de $\mathbf{Z}^r$. Une condition nécessaire et suffisante de résolution est que tous les déterminants d'ordre r extraits de la matrice des coordonnées de (C, V,, V_2, ... V,,) soient divisibles par le P.G.C.D. des déterminants d'ordre Y extraits de la matrice des coordonnées de (V,, V_2, ... V,,).

Signalons que le *théorème des restes chinois* ($x \equiv a_i$ mod m_i, pour $i = 1, 2, \ldots$ r) correspond à un cas non homogène, avec $n = r + 1$. Ii se ramène, si les m_i sont premiers deux à deux, à une seule équation : $x \equiv$ a mod m, avec $m = m_1 m_2 m_r$.

Généralités sur le second degré

La résolution en entiers de :

$$ax^2 + bxy + cy^2 + dx + ey + k = 0,$$

équation de conique à coefficients entiers, n'est intéressante que dans les cas parabolique ou hyperbolique. L'étude en a été faite par Euler et Lagrange. Dans le cas elliptique, en effet, il n'y a qu'un nombre fini (éventuellement nul) de solutions, qu'on peut déterminer par essais successifs. C'est ainsi que Gauss a étudié l'équation $ax^2 + by^2 = m$, avec a et b entiers positifs.

Dans le cas parabolique, on pose 2 $ax + by = t$, d'où 4 $adx + 4\,aey = t^2$ 4 ak et la résolution, lorsqu'elle est possible, conduit à des formules du type :

$$\begin{cases} x = x_0 + u_1\lambda + v_1\lambda^2 \\ y = y_0 + u_2\lambda + v_2\lambda^2, \end{cases}$$

ou $x_0, y_0, u_1, v_1, u_2, v_2$ sont des entiers fixés et où A parcourt Z. De telles formules nécessitent, en général, diverses constantes pour donner toutes les solutions.

Par exemple : 4 $x^2 - 20\,xy + 25\,y^2 + 14\,x - 41\,y + 18 = 0$ conduit aux quatre systèmes de solutions :

$$\begin{cases} x = 2 - 19\lambda + 60\lambda^2 \\ y = 2 - 10\lambda + 24\lambda^2 \end{cases}$$
$$\begin{cases} x = 1 + 11\,A + 60\lambda^2 \\ y = 1 + 2\lambda + 24\lambda^2 \end{cases}$$
$$\begin{cases} x = 16 + 61\,A + 60\lambda^2 \\ y = 6 + 22\lambda + 24\lambda^2 \end{cases}$$
$$\begin{cases} x = 35 + 91\lambda + 60\lambda^2 \\ y = 13 + 34\lambda + 24\lambda^2. \end{cases}$$

À l'inverse, 2 $x^2 + 2\,y + 1 = 0$ n'a évidemment pas de solution.

Dans le cas hyperbolique, on se ramène au centre de coordonnées rationnelles $\alpha = p/\Delta$, $\beta = q/\Delta$, et on posc $x = (p + X)/\Delta$, $y = (q + Y)/\Delta$, qui conduit à :

$$aX^2 + bXY + cY^2 = m.$$

Le cas $b^2 - 4\,ac = D$ carré parfait fournit :

$$(u_1X + v_1Y)(u_2X + v_2Y) = m,$$

d'où la résolution par un nombre fini de systèmes linéaires :

$$\begin{cases} u_1X + v_1Y = m_1 \\ u_2X + v_2Y = m_2. \end{cases}$$

Si $b^2 - 4\,ac = D$ n'est pas carré parfait, on est amené à faire intervenir les solutions de :

$$u^2 + buv + acv^2 = \pm 1$$

(liées aux unités de Q ($\sqrt{d}$), comme on le voit à propos des équations de Pell). On obtient alors les solutions de :

$$aX^2 + bXY + cY^2 = m,$$

à partir d'un nombre fini d'entre elles. II peut d'ailleurs n'y avoir aucune solu-

tion, comme, par exemple, pour $x^2 - 3y^2 = -1$.

L'équation $3x^2 \quad 2xy \quad 2y^2 = 6$, elle, conduit à deux séries de solutions, données par :

$$3x_n - \gamma y_n = (6-\gamma)(11-3\gamma)^n$$

et $$3x'_n - \gamma y'_n = (12-3\gamma)(11-3\gamma)^n,$$

où $y = 1 + \sqrt{7}$ est racine de $\gamma^2 \quad 2y \quad 6 = 0$ et où $(11 - 3y)$” correspond aux solutions de $u^2 - 2uv - 6v^2 = 1$.

Équation de Pell

L'équation de Pell :

$$x^2 - dy^2 = m$$

(appelée également équation de Pell-Fermat), où l'on suppose d sans facteur carré, joue un rôle particulier dans les équations du second degré ; elle est, en effet, fondamentalement liée à la recherche des unités du corps quadratique Q $(\sqrt{d})$. Après avoir établi que $x^2 \quad dy^2 = m$ a une infinité de solutions entières pour au moins un m tel que $|m < 1 + 2\sqrt{d}$, on classe ces solutions modulo $m|$, d'où l'existence des solutions de $x^2 - dy^2 = 1$.

Si l'on associe **aux** solutions de $x^2 - dy^2 = 1$ les nombres $\xi = x + y\sqrt{d}$, on voit qu'ils forment un groupe multiplicatif et on établit l'existence d'une solution fondamentale (x_1, y_1) telle que toutes les solutions de $x^2 - dy^2 = 1$ sont données par :

$$x + y\sqrt{d} = \pm(x_1 + y_1\sqrt{d})^n,$$

avec $n \in \mathbf{Z}$.

Pour $m \neq 1$, l'équation peut ne pas avoir de solution ($x^2 \quad 3y^2 = -1 + 4k$ par exemple, car $x^2 \quad 3y^2$ est congru à 0.1 ou 2, modulo 4). Si $x^2 - dy^2 = -1$ est résolue en entiers, ses solutions seront données à partir d'une solution fondamentale (x_0, y_0) par :

$$x + y\sqrt{d} = \pm(x_0 + y_0\sqrt{d})^{2n+1},$$

où $n \in \mathbf{Z}$ et $(x_0 + y_0\sqrt{d})^2 = x_1 + y_1\sqrt{d}$ défini ci-dessus.

Le cas général $x^2 \quad dy^2 = m$, lorsqu'il aura des solutions, permettra de ranger celles-ci en un certain nombre s de classes, données par :

$$x + y\sqrt{d} = \pm(u_k + v_k\sqrt{d})(x_1 + y_1\sqrt{d})^n,$$

où $n \in \mathbf{Z}$ et $k = 1, 2 \ldots, s$.

Le cas $m = \pm 4$ est plus directement lié aux unités de Q $(\sqrt{d})$, qui se recherchent sous la forme $(x + y\sqrt{d})/2$, solution de $\xi^2 + p\xi \pm 1 = 0$ avec $p^2 \pm 4 = u^2 d$. Il y a évidemment toujours des solutions pour $m = +4$, données par :

$$\pm 2\left(\frac{x'_0 + y'_0\sqrt{d}}{2}\right)^n,$$

où $n \in \mathbf{Z}$, avec (x'_0, y'_0) plus petite solution positive. Pour $m = \quad 4$, lorsqu'il y a solutions, elles s'exprimeront de même sous une forme :

$$\pm 2\left(\frac{x''_0 + y''_0\sqrt{d}}{2}\right)^{2n+1}, \; n \in \mathbf{Z}.$$

L'équation de Pell est aussi liée aux développements en fraction continuée de $\sqrt{d}$. En effet, on établit que les solutions de $x^2 - dy^2 = 1$ correspondent à des réduites p_n/q_n du développement de $\sqrt{d}$. Plus précisément, l'unité fondamentale de Q $(\sqrt{d})$ est $(x_0 = y_0\sqrt{d})$, avec les notations précédentes, si le développement de $\sqrt{d}$ est de période impaire (il y a alors des solutions à $x^2 - dy^2 = -1$), alors que c'est $(x_1 + y_1\sqrt{d})$, si la période est paire.

Dans les deux cas, le développement de :

$$\frac{1}{2}[\mathrm{I}(\sqrt{d}) + \sqrt{d}],$$

où $\mathrm{I}(\sqrt{d})$ est le plus grand nombre impair inférieur à $\sqrt{d}$, fournit cette solution fondamentale par sa k-ième réduite, si la période est ***k.***

Coniques

L'équation :

$$ax^2 + by^2 + cz^2 = 0,$$

où l'on peut supposer a, b et c sans facteurs carrés et premiers entre eux deux à deux, conduit à un théorème de Legendre : une condition nécessaire et suffisante de résolubilité est que a, b, c ne soient pas de même signe et que bc, $-$ CU et $-$ ah soient respectivement résidus quadratiques de a, b et c (un résidu quadratique de α est un entier β premier avec α tel que $x^2 \equiv \beta$ mod α soit résoluble en x ; cf. DIVISIBILITÉ, chap. 4), et alors il y a une solution avec $x \leqslant \sqrt{2|bc|}$, /y $\leqslant \sqrt{2ca}$, $z \leqslant \sqrt{2|ab|}$. On a d'ailleurs là un cas particulier du théorème de Minkowski-Hasse, suivant lequel : une forme quadratique à coefficients rationnels représente zéro (c'est-à-dire s'annule pour un système de valeurs non toutes nulles des variables dans le corps considéré) dans le corps des rationnels si, et seulement si, elle représente zéro non trivialement dans tous les corps p-adiques et dans le corps des réels. C'est dire encore qu'une forme quadratique à coefficients entiers ne s'annulera pour des valeurs entières des variables que si, et seulement si, son égalité à zéro, modulo p^m, est résoluble, quels que soient le nombre premier p et l'entier naturel m, en solutions entières non toutes divisibles par p. On dit alors que les conditions de congruence sont satisfaites,

Ici encore on sait majorer la taille d'une plus petite solution. Un théorème de Chevalley, sur les formes de degré strictement inférieur au nombre des variables, permet d'affirmer que :

$$f(x_1, x_2, ..., x_n) = 0 \pmod p$$

a une solution non nulle, si $n \geqslant 3$, pour toute forme quadratique. En combinant cela avec le lemme de Hensel, on voit que les conditions de congruence à vérifier sont en nombre fini.

Dans le cas de deux variables :

$$ax^2 + 2bxy + cy^2 = 0 \pmod p,$$

avec $p \neq 2$, a une solution et une seule si, et seulement si, $(ac - b^2)$ est ou bien un multiple de p, ou bien l'opposé d'un résidu quadratique de p.

Notons enfin que Dickson a établi que, si :

$$ax^2 + by^2 + cz^2 = 0$$

est résoluble, avec a, b, c premiers entre eux deux à deux et sans facteurs carrés, il s'ensuit que tout entier peut s'exprimer sous la forme $ax^2 + by^2 + cz^2$.

Équation de Pythagore

L'équation de Pythagore :

$$x^2 + y^2 = z^2,$$

est un cas suffisamment classique pour qu'on s'y arrête. Si l'on suppose (x, y, z) solution en entiers premiers entre eux, l'un des deux nombres x ou y est pair. Supposons que ce soit x. 11 vient :

$$\left(\frac{x}{2}\right)^2 = \left(\frac{z+y}{2}\right)\left(\frac{z-y}{2}\right)$$

qui entraîne :

$$(z+y)/2 = a^2 \quad \text{et} \quad (z-y)/2 = b^2,$$

d'où les solutions générales positives données par :

$$x = 2ab,\ y = a^2 - b^2,\ z = a^2 + b^2,$$

avec (a, b) = 1. $a > b > 0$ et a et b de parité différente. On trouve pour a = 2 et b = 1 la plus petite solution non banale : (4, 3, 5).

Il est intéressant de signaler que l'étude du groupe orthogonal de la forme qua-

dratique $(x^2 + y^2 \quad z^2)$ conduit à envisager ici les matrices de ce groupe, à éléments entiers. On peut établir que, si :

$$S = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \end{pmatrix},$$

$$T = \begin{pmatrix} -1 & 2 & 2 \\ -2 & 1 & 2 \\ -2 & 2 & 3 \end{pmatrix},$$

$$U = \begin{pmatrix} 1 & -2 & 2 \\ 2 & -1 & 2 \\ 2 & -2 & 3 \end{pmatrix},$$

toutes les solutions entières, premières entre elles, de l'équation de Pythagore, peuvent s'obtenir à partir de la solution (3, 4, 5) par application répétée de S, T, U dans un ordre quelconque (cf. figure).

Avant d'abandonner le second degré, citons l'équation :

$$x^2 + y^2 = z^2 + t^2,$$

et, plus généralement, le système :

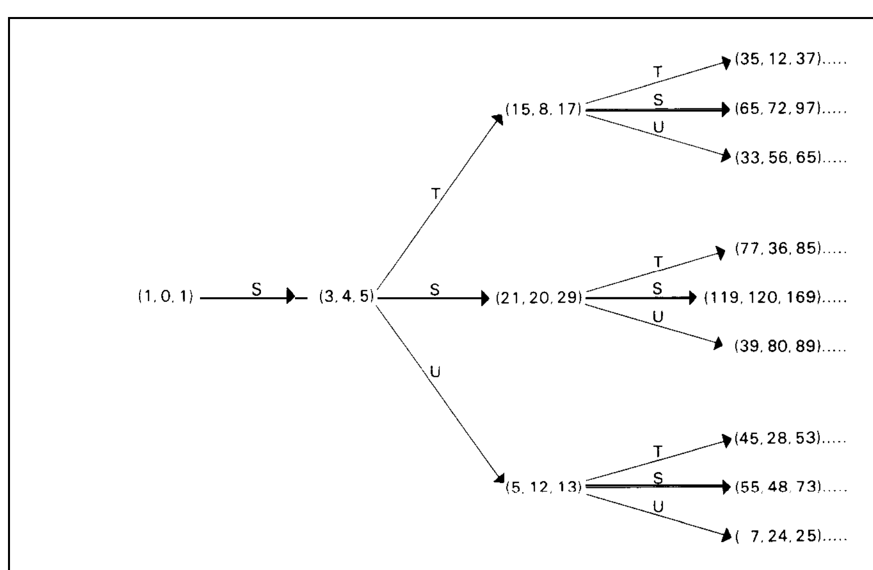
$$x_1^2 + y_1^2 = x_2^2 + y_2^2 = \dots = x_r^2 + y_r^2,$$

en liaison avec l'étude du nombre des représentations p,(n) d'un nombre n comme somme de deux carrés ; on démontre que :

$$\lim_{n \to +\infty} \rho_2(n) = +\infty.$$

2. Le grand théorème de Fermat

Pierre de Fermat (1601-1665) fut un mathématicien d'une érudition extraordinaire (géométrie analytique, fondements du calcul infinitésimal, lois de l'optique, fondements du calcul des probabilités et surtout théorie des nombres). Malheureusement, presque tous ses théorèmes étaient donnés sans démonstration, car il était alors d'usage de proposer ses découvertes à la sagacité de ses interlocuteurs (avec en particulier une rivalité très vive entre géomètres anglais et géomètres français). Le théorème élémentaire de Fermat



Solutions entières premieres entre elles de l'équation de Pythagore

($a^p - a$ est toujours divisible par p si p est premier), de même que toutes ses études sur les formes quadratiques et sur l'équation de Pell (appelée souvent Pell-Fermat) ont été vérifiés et établis dès le XVIII[e] siècle, ainsi que la plupart des énoncés qu'il a affirmés, à l'exception de ce que l'on appelle le *grand théorème de Fermat* (on dit aussi le « dernier théorème de Fermat »). La recherche d'une démonstration de ce résultat a constitué, comme on le voit ci-dessous, une motivation essentielle dans le développement des mathématiques et a contribué à l'élaboration de l'algèbre moderne. Une démonstration définitive a été donnée par le mathématicien britannique A. Wiles en 1993-1995.

Ce « théorème » consiste en la proposition suivante : Pour $n \geqslant 3$, l'équation

$$x^n + y^n = z^n$$

est impossible en nombres entiers avec $xyz \neq 0$. L'auteur affirme cette proposition, en 1637, dans une annotation marginale des œuvres de Diophante ; il y écrit : « J'ai découvert une démonstration assez remarquable de cette proposition, mais elle ne tiendrait pas dans cette marge. »

Il suffirait d'établir ce théorème pour $n = 4$ et pour tout nombre premier p. Malheureusement, si la démonstration pour $x^4 + y^4 = z^4$ est assez simple par la méthode de descente infinie, si Euler et Gauss traitent le cas de y = 3 de même, si Legendre en 1823 met au point le cas de $p = 5$ (par montée infinie), le théorème a été très difficile à établir dans sa généralité.

Pour $n = 4$, la démonstration de Frénicle (1676) repose sur la descente infinie dont le principe était donné par Fermat : raisonnant sur l'équation de Pythagore $(x^2)^2 + (y^2)^2 = z^2$, on obtient, à partir de toute solution éventuelle (x_0, y_0, z_0), une nouvelle solution où $z_1 < |z_0|$, ce qui permet de conclure à l'impossibilité.

Notons aussi l'impossibilité de $x^4 + y^4 = 2z^2$, si $x^2 \neq y^2$ (d'où l'impossibilité de trouver trois entiers dont les puissances quatrièmes soient en progression arithmétique de raison non nulle).

De même $x^4 + y^4 = 3z^2$ est impossible (comme $x^2 + y^2 = 3z^2$) et, d'une manière plus générale, $x^4 + y^4 = kz^2$ est impossible pour $3 \leqslant k \leqslant 16$, sauf $k = 8$ (études générales de Maillet en 1900).

Pour $n = 3$, la démonstration ébauchée par Euler en 1774 fut précisée par Gauss. Comme dans le cas général de n premier quelconque, la recherche se scinde en deux étapes : on montre d'abord l'impossibilité en nombres non divisibles par 3 ; pour cela, on déduit de $x^3 + y^3 = z^3$ la congruence :

$$x^3 + y^3 \equiv z^3 \pmod 9,$$

d'où :

$$x + y \equiv z \pmod 3, \quad \text{car } a \equiv a^3 \pmod 3,$$

d'où :

$$x^3 + y^3 = (x + y + 3u)^3 \equiv x^3 + y^3 + 3xy(x + y) \pmod 9,$$

donc : $xy(x + y) \equiv xyz \equiv 0 \pmod 3,$

ce qui est impossible sans que x, y ou z soit divisible par 3.

La dernière étape est plus délicate et repose sur une descente infinie : on suppose une solution de $x^3 + y^3 = z^3$, avec xyz minimum et on pose :

$$x = u + w, \quad y = u - w,$$

d'où : $$2u(u^2 + 3w^2) = z^3,$$

qui conduit à $u^2 + 3w^2 = s^3$, si $(u, 3) = 1$. Si, au contraire, 3 divise u, on pose $u = 3v$, d'où $18v(3v^2 + w^2) = z^3$ conduisant à $3v^2 + w^2 = s^3$. Les solu-

tions de $s^3 = a^2 + 3\,b^2$, mises sous la forme :

$$s = \alpha^2 + 3\beta^2, \quad a = \alpha^3 - 9\alpha\beta^2$$
$$\text{et } b = 3\alpha^2\beta - 3\beta^3,$$

conduisent alors à :

$$2a = \sigma^3, \quad \alpha - 3\beta = \tau^3 \quad \text{et} \quad a + 3\beta = \rho^3,$$

où $\rho^3 + \tau^3 = \sigma^3$, avec $\rho\sigma\tau < xyz\,|$.

Pour $n = 5$, la démonstration faite par Legendre en 1825 repose sur :

$$x^5 = (z - y)\,\varphi(x, y),$$
$$\text{où } 4\varphi(x,y) = 5(x^2 + y^2)^2 - (x - y)^4$$
$$= (2x^2 + xy + 2y^2)^2 - 5(xy)^2\,;$$

cela permet d'établir que $\varphi(x, y)$ doit avoir des diviseurs aussi grands qu'on veut (méthode de montée infinie).

La mathématicienne Sophie Germain a établi que, si n est premier ainsi que $(2n + 1)$, il faudrait, pour que l'équation de Fermat soit vérifiée, que x, y ou z soit divisible par n. Ce résultat a été généralisé par Legendre.

Lamé, en 1837, établit le cas $n = 7$ après que Lejeune-Dirichlet ait démontré, en 1832, l'impossibilité pour $n = 14$.

L'étude **générale,** pour n premier impair, comprend donc deux étapes. Dans la première étape, appelée souvent « premier cas du théorème », on démontre qu'il n'y a pas de solution parmi les entiers non multiples de n. Dans la deuxième étape. appelée « deuxième cas du théorème », on montre qu'il n'y a pas de solution dont l'un dcs nombrcs soit multiple de n. Cette étude générale fut entreprise par Kummer en 1844 et utilise le corps Q (p) des nombres algébriques de degré $(n - 1)$ définis par l'équation $\rho^n - 1 = 0$. En cffct, si α est une racine primitive n-ième de l'unité, l'équation de Fermat s'écrit :

$$x^n = (z - y)\,\varphi(z, y),$$

avec :

$$\varphi(z, y) = (z + y\alpha)(z + y\alpha^2)\ldots(z + y\alpha^{n-1}).$$

C'est en se plaçant dans Q (p) que Kummer essaya, à partir des entiers complexes $m_0 + m_1\alpha + \ldots + m_{n-1}\alpha^{n-1}$ (m, entiers), de raisonner par décomposition en facteurs premiers de $\varphi(z, y)$. Malheureusement, cette décomposition n'est pas toujours unique, et c'est a cette occasion que Kummer introduit la notion de nombre *idéal* (Cf. ANNEAUX COMMUTATIFS). Ces nombres idéaux, n'appartenant pas au corps envisagé, permettent la décomposition unique en facteurs idéaux premiers. Cette notion d'idéal fut précisée, un peu plus tard, d'un point de vue purement algébrique, par Dedekind.

Kummer obtint des résultats spectaculaires, mais encore incomplets : le théorème de Fermat est vérifié pour **tout** premier p pour lesquels le nombre de classes d'idéaux n'est pas divisible par p (un tel nombre p est appelé nombre premier régulier). On ne sait pas, actuellement, s'il existe un nombre infini de nombres premiers réguliers ; seuls 37, 59 et 67 sont non réguliers dans la première centaine. Mirimanoff en 1893 démontre le cas $p = 37$ en perfectionnant la méthode de Kummer. En 1968, on avait établi le théorème de Fermat pour tous les nombres premiers jusqu'à 125 000, et pour un certain nombre d'autres. La démonstration de Wiles clôt une longue histoire de tentatives infructueuses.

3. Méthodes géométriques

Pour classer les types d'équations, on utilise d'abord la dimension, ou nombre de

variables indépendantes, du système proposé. Ainsi, en général, un système :

$$f_1(x_1, ..., x_n) = 0$$
$$. . . . .$$
$$f_r(x_1, ..., x_n) = 0$$

est de dimension $(n - r)$. En dimension 1, on parle de courbes ; en dimension 2, de surfaces (cf. GÉOMÉTRIE ALGÉBRIQUE). Les solutions en nombres entiers ou rationnels du système proposé ne sont autres que les points entiers ou rationnels de la variété algébrique associée.

Déjà pour les courbes planes (une équation $f(x, y) = 0$), la classification par le degré s'avère trop grossière. Ainsi la théorie des cubiques planes à point double, comme :

$$y^2 = x^2(x + a),$$

avec a rationnel, est très simple, celle des cubiques planes sans point double, comme :

$$y^2 = x(x + a)(x + b),$$

avec a et b rationnels non nuls et distincts, est beaucoup plus délicate (cf. COURBES ALGÉBRIQUES). On est ainsi amené à utiliser des « invariants » de nature géométrique. Pour les courbes, on utilise le genre, nombre de trous de la surface de Riemann correspondante. Pour une courbe plane de degré d, à points multiples ordinaires, le genre vaut :

$$\frac{(d-1)(d-2)}{2} \sum \frac{s_i(s_i - 1)}{2},$$

où la sommation s'étend à tous les points multiples, l'ordre de ceux-ci étant s_i.

Courbes de genre zéro

On dispose ici d'une analyse complète, En ce qui concerne les points rationnels, la démarche est la suivante. Toute courbe de genre zéro peut. par un changement de variables, être ramenée à une conique plane (D. Hilbert-A. Hurwitz, 1891), soit $ax^2 + by^2 + c = 0$. D'après le théorème de Legendre (cf. *supra*), les conditions de congruence permettent de décider si cette conique a un point rationnel. S'il y a un point rationnel, soit M_0, on peut les décrire tous, au moyen d'une paramétrisation biunivoque :

$$x = f(t), Y = g(t),$$

avec f et g des quotients de polynômes à coefficients rationnels (chaque point rationnel de la conique correspond à une unique valeur, rationnelle, du paramètre t). Une telle paramétrisation sera dite polynomiale. Pour obtenir cette paramétrisation, on fait simplement tourner une droite autour de M_0 ; chaque droite de pente rationnelle t, passant par M_0, recoupe la conique, qui est de degré 2, en un unique point, à coordonnées $(x = f(t), y = g(r))$ rationnelles. C'est ainsi qu'a été résolue plus haut l'équation de Pythagore.

Pour les points entiers sur une courbe de genre zéro, on dispose aussi d'une analyse complète (C. L. Siegel, 1929) : essentiellement, l'équation de Pell (cf. *supra*) est le seul cas non évident où il peut y avoir une infinité de points entiers.

Courbes de genre 1 : points rationnels

Ici, les conditions de congruence ne suffisent plus à assurer l'existence d'un point rationnel, comme le montre l'exemple $3x^3 + 4y^3 + 5 = 0$ (E. S. Selmer, 1951). On dispose cependant d'un procédé remontant à Fermat (descente infinie) permettant d'étudier de telles courbes. C'est un problème ouvert de savoir si l'application systématique de ce procédé, conjointement avec les conditions de

congruence, suffit toujours à déterminer la présence ou l'absence d'un point rationnel sur une courbe de genre 1.

Si l'on connaît un point rationnel sur une telle courbe, celle-ci peut être ramenée (Poincaré, 1901) à une cubique plane non singulière :

$$(C) \qquad y^2 = P(x),$$

avec $P(x)$ un polynôme du troisième degré sans facteur multiple. On a là une courbe elliptique, objet fondamental tant en géométrie algébrique qu'en théorie des nombres. La géométrie algébrique montre qu'on ne peut pas ramener une telle courbe à une courbe de genre zéro. Il existe pourtant une paramétrisation des solutions complexes de (C), bien classique, mais, elle, de nature transcendante : la paramétrisation par les fonctions elliptiques (Weierstrass; cf. COURBES ALGÉBRIQUES). On voit sur cette paramétrisation que les solutions complexes de (C) peuvent être munies d'une structure de groupe abélien. En fait, la composition ainsi définie induit une composition des points rationnels de la cubique. Essentiellement, il s'agit de la construction suivante. Étant donné deux points rationnels de la cubique, la droite qui les joint, qui est à coefficients rationnels, recoupe la cubique en un troisième point, dont les coordonnées sont par force rationnelles : c'est le procédé de la *corde* pour engendrer de nouvelles solutions (on peut aussi utiliser la *tangente* en un point rationnel). On obtient ainsi une structure de groupe abélien sur l'ensemble des points rationnels de (C). L'important théorème de Mordell (1922), généralisé par Weil (1928), établi par descente infinie, dit que ce groupe appelé depuis groupe de Mordell-Weil, admet un nombre fini de générateurs. En d'autres termes, étant donné une cubique (C), il existe un nombre fini de points rationnels situés dessus tels que tous les autres points rationnels de (C) puissent être obtenus à partir de ceux-ci par itération du procédé de la corde et de la tangente. Étant donné une courbe (C), on sait borner le nombre minimal de générateurs du groupe de Mordell-Weil associé *(rang)*, mais on n'a pas de méthode générale pour déterminer le rang, a fortiori pour construire explicitement un système de générateurs. On dispose seulement d'un algorithme conditionnel (Yu. I. Manin, 1973), reposant sur deux conjectures. L'une, de Weil, relie les courbes elliptiques aux formes modulaires. L'autre, de B. Birch et H. P. F. Swinnerton-Dyer (1965), affirme que le rang d'une courbe elliptique est donné par l'ordre du zéro au point complexe $z = 1$ d'une certaine fonction méromorphe dont la définition fait intervenir le nombre de solutions modulo p, nombre premier, de l'équation (C), cela pour tous les nombres premiers p (cf. fonction ZÊTA). L'ordinateur a donné un grand poids à cette conjecture, et des progrès théoriques ont été effectués.

Courbes de genre au moins égal à 2 : points rationnels

Parmi les courbes de genre au moins égal à 2, on trouve les courbes planes non singulières de degré au moins 4. Là encore. on ne peut ramener l'étude de telles courbes à l'étude de celles de genre inférieur ou égal à 1. On ne dispose d'aucun procédé permettant d'engendrer un nombre infini de solutions à partir d'un nombre fini d'entre elles. Mordell a ainsi conjecturé (1922) qu'une telle équation n'admettrait jamais qu'un nombre fini de solutions rationnelles. Cette conjecture a été démontrée par G. Faldings en juin 1983 et a constitué une étape importante

dans la démonstration du grand théorème de Fermat.

Points entiers sur les courbes de genre au moins 1

On dispose du théorème général de C. L. Siegel (1929) selon lequel une telle courbe n'a qu'un nombre fini de points entiers. La démonstration utilise d'une part le théorème de A. Weil (1928), étendant celui de Mordell, d'autre part la mauvaise approximation par des rationnels des irrationnels algébriques (cf. approximations DIOPHANTIENNES). Ce résultat englobe celui de Thue (1909), lui aussi fondé sur les approximations diophantiennes : l'équation

$$f(x, Y) = m,$$

où f est un polynôme homogène irréductible à coefficients entiers de degré au moins égal à 3, et m est un entier non nul, ne possède qu'un nombre fini de solutions entières.

Dans beaucoup de cas, en particulier pour les équations $y^2 = P(x)$ où le polynôme P a au moins trois zéros distincts, A. Baker a donné des majorations effectives (mais grandes) pour la taille possible des solutions entières ; ainsi, pour l'équation de Thue ci-dessus :

$$\max(|x|, |y|) < \exp\{(d\,H)^{(10d)^5}\},$$

où d est le degré de f, et H un entier dépendant de la taille de m et des coefficients de f. Dans certains cas particuliers, ces méthodes permettent même de trouver toutes les solutions entières,

Surfaces rationnelles

Les surfaces rationnelles sont les analogues en dimension 2 des courbes unicursales, celles qui peuvent être paramétrées de façon polynomiale si l'on autorise les coefficients des polynômes définissant le paramétrage à être des nombres complexes. Parmi celle-ci, on trouve les surfaces non singulières de l'espace ordinaire définies par une équation de degré 2 (quadriques) ou 3 (surfaces cubiques), mais aussi des équations de degré supérieur, comme :

$$y^2 + a(x)z^2 + b(x) = 0,$$

avec $a(x)$ et $b(x)$ des polynômes non nuls de degré quelconque.

On est loin de disposer ici de résultats aussi satisfaisants que pour les courbes de genre zéro. Dans certains cas : quadriques (Hasse-Minkowski, cf. *supra),* surfaces cubiques

$$ax^3 + by^3 + cz^3 + d = 0,$$

avec $ab = cd \neq 0$ (E. S. Selmer, 1953), les conditions de congruence (et la condition réelle) suffisent à assurer l'existence de points rationnels. Cela ne vaut pas en général, comme le montre l'exemple (J. W. S. Cassels-M. J.T. Guy, 1966) :

$$5x^3 + 9y^3 + 10z^3 + 12 = 0$$

Si une quadrique a un point rationnel, on peut encore donner une paramétrisation polynomiale essentiellement biunivoque des points rationnels, en utilisant la même méthode que pour les coniques.

Une telle paramétrisation est encore possible pour une surface cubique de l'espace ordinaire, soit Σ, qui contient deux droites D et D' définies par des équations à coefficients rationnels, et ne se coupant pas. Choisissons un plan π, d'équation $ax + by + cz + d = 0$, avec a, b, c, d rationnels. Pour M un point rationnel du plan π, la droite D_M, intersection des deux plans engendrés l'un par (M, D), l'autre par (M, D'), est définie par des équations

à coefficients rationnels, Elle coupe la surface cubique Σ en trois points : celui qui est situé sur D et celui qui est situé sur D' sont définis par des équations à coefficients rationnels, le troisième, soit $f(M)$, est donc à coordonnées rationnelles. On vérifie que la correspondance qui à M associe J'(M) définit une paramétrisation poiynomiale essentiellement biunivoque des points rationnels de Σ par ceux de π. C'est ainsi qu'on trouve la solution générale due à Euler de l'équation : $x^3 + y^3 + z^3 = 1$

$$x = \frac{1 - (u - 3v)(u^2 + 3v^2)}{d}$$

$$y = \frac{(u + 3v)(u^2 + 3v^2) - 1}{d}$$

$$z = \frac{(u^2 + 3v^2)^2 - (u + 3v)}{d}$$

avec $d = (u^2 + 3v^2)^2 - (u - 3v)$.

Des méthodes fines de géométrie algébrique montrent qu'une telle paramétrisation polynomiale biunivoque à coefficients rationnels est souvent impossible, ainsi pour :

$$x^3 + y^3 + z^3 = a,$$

où a est un rationnel qui n'est pas un cube (Yu. 1. Manin, 1970). On peut cependant, pour une surface cubique non singulière qui possède au moins un point rationnel, trouver des familles polynomiales à deux vrais paramètres (c'est-a-dire qu'on ne peut réduire à un seul paramètre), soit :

$$x = A(u, v),\ y = B(u, v),\quad z = C(u, v),$$

avec A,B,C quotients de polynômes en u et y à coefficients rationnels, mais en général même un nombre fini de telles familles ne suffit pas à décrire tous les points rationnels de la surface cubique. On obtient néanmoins beaucoup de solutions rationnelles En spécialisant les paramètres, on obtient par exemple que tout rationnel est somme de trois cubes de rationnels :

$$a = \left(\frac{a^3 - 3^6}{d}\right)^3 + \left(\frac{-a^3 + 3^5 a + 3^6}{d}\right)^3, + \left(\frac{a^2 + 3^4 a}{d}\right)^3,$$

avec $d = 3^2 a^2 + 3^4 a + 3^6$

(Ryley, 1825 ; H. W. Richmond, 1930).

En s'inspirant de la méthode de Mordell-Weil, F. Châtelet (1959) a montré qu'un nombre fini de solutions paramétriques polynomiales (à 4 variables) permet de décrire toutes les solutions rationnelles de :

$$y^2 - az^2 = x(x - b)(x - c),$$

avec a, b, c rationnels non nuls. On ignore par contre si une solution polynomiale, même partielle, à deux vrais paramètres et à coefficients rationnels est possible pour l'équation générale :

$$y^2 - a(x)z^2 = b(x),$$

où $a(x)$ et b(x) sont des polynômes non nuls.

Pour les points *entiers* des surfaces cubiques, on a des résultats épars. Soit par exemple l'équation :

$$x^3 + y^3 + z^3 = n,$$

avec n entier fixé, à résoudre en (x, y, z) entiers. On voit facilement (congruences modulo 9) qu'il n'y a pas de solutions si II est congru à ± 4 modulo 9. Sinon, on ne sait pas s'il y a toujours une solution, par exemple pour $n = 30$, et, s'il y en a une, s'il y en a une infinité. L'identité :

$$(1 + 6t^3)^3 + (1 - 6t^3)^3 + (-6t^2)^3 = 2$$

donne une infinité de solutions pour $n = 2$, mais, pour $n = 3$, on ne connaît que les solutions (1, 1, 1) et (4, 4, 5).

Pour :

$$x^2 + y^2 + z^2 - axyz = b,$$

avec a et b entiers, on dispose d'un processus (Markov, Hurwitz) permettant d'engendrer toutes les solutions entières à partir d'un nombre fini, explicitement calculables, d'entre elles.

On conjecture (P. Erdös, E. G. Straus) que l'équation :

$$4/n = 1/x + 1/y + 1/z,$$

avec n entier fixé, $n > 1$, à résoudre en entiers positifs x, y, z, a toujours des solutions (c'est un cas particulier du problème des *fractions égyptiennes, où* l'on cherche à écrire un rationnel comme somme d'un nombre donné d'inverses d'entiers). Cette conjecture a été établie pour $n \leqslant 10^8$.

Surfaces analogues aux courbes de genre 1

L'analogue immédiat, du point de vue de la géométrie algébrique, consiste en les surfaces abéliennes (pour lesquelles il est difficile de donner des équations !). Le théorème de A. Weil (1928) nous renseigne sur les points rationnels, mais on ignore s'il n'y a qu'un nombre fini de points entiers.

Un autre analogue consiste en les surfaces non singulières de l'espace ordinaire, de degré 4. Une conjecture d'Euler affirme que l'équation :

$$x^4 + y^4 + z^4 = 1$$

n'a pas d'autres solutions rationnelles que $(\pm 1, 0, 0)$, $(0, \pm 1, 0)$ et $(0, 0, \pm 1)$, mais on sait seulement qu'une autre solution devrait avoir un dénominateur au moins égal à 220 000. Euler donna une infinité de solutions rationnelles de :

$$x^4 + y^4 = z^4 + 1,$$

au moyen d'une solution à un paramètre ; par exemple :

$$x = \frac{t^7 + t^5 - 2t^3 + 3t^2 + t}{d}$$

$$y = \frac{t^6 - 3t^5 - 2t^4 + t^2 + 1}{d}$$

$$z = \frac{t^7 + t^5 - 2t^3 - 3t^2 + t}{d}$$

avec $d = t^6 + 3t^5 - 2t^4 + t^2 + 1$ (Géraudin, 1917).

La géométrie algébrique montre qu'aucune solution, même partielle, polynomiale à deux vrais paramètres n'est possible. Cependant, en utilisant le fait que la surface ci-dessus peut être fibrée en courbes elliptiques, on peut donner une infinité de solutions essentiellement distinctes à un paramètre (H. P. F. Swinnerton-Dyer, 1971).

Surfaces analogues aux courbes de genre au moins 2

Parmi celles-ci, on trouve les surfaces non singulières d'équation :

$$f(x, y, z) = 0,$$

avec f de degré au moins 5. Une fois de plus, elles ne peuvent être ramenées aux surfaces précédentes. Il peut exister des solutions à un paramètre, mais une conjecture analogue à celle de Mordell pour les courbes de genre au moins 2 voudrait que, pour une telle surface, les points rationnels soient concentrés sur un nombre fini de courbes algébriques tracées sur la surface.

4. Équations à beaucoup de variables

La méthode du cercle de Hardy-Littlewood-Vinogradov, qui avait déjà révélé sa puissance dans l'étude du problème de Waring (cf. théorie des NOMBRES - Théorie analytique des nombres), a aussi

permis d'obtenir le résultat suivant (H. Davenport, B. Birch, 1962). Soit $f_1, \ldots, f_r$ des formes homogènes de degré d, en n variables, à coefficients entiers. Supposons que le système :

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ \ldots\ldots\ldots\ldots \\ f_r(x_1, \ldots, x_n) = 0 \end{cases}$$

n'a pas de solution complexe singulière non nulle. Si les conditions de congruence sont satisfaites, si le système a une solution non nulle en nombres réels et si :

$$(*) \qquad n > r(r+1)\,(d-1)\,2^{d-1},$$

alors le système a une solution non nulle en nombres entiers, On peut appliquer ce résultat par exemple à une forme cubique en au moins 17 variables (toute forme telle a un zéro non trivial), ou à un système de deux formes quadratiques en au moins 13 variables.

Si les conditions de congruence (et la condition réelle) ne sont pas en général suffisantes quand le nombre de variables est très petit par rapport au degré (ainsi $r = 1$, $d = 3$, $n = 4$, cf. ***supra),*** on peut se demander si on ne peut pas affaiblir l'inégalité $(*)$. On ignore ainsi si, pour une forme cubique non singulière en **au** moins 5 variables, les conditions de congruence sont suffisantes.

On se demande si une équation non singulière :

$$f(x_1, \ldots, x_n) = 0$$

de degré $d \leqslant n$ a une infinité de solutions rationnelles dès qu'elle en a une. Considérons par exemple l'équation :

$$x_1^d + \ldots + x_n^d = 1.$$

Pour $d = n = 4$, à part les solutions évidentes du type (1, 0, ... 0), on connaît :

(240/651, 340/651, 430/651, 599/651),

mais on ne connaît pas de solution à un paramètre. Pour $d = n = 5$, on connaît des solutions à deux paramètres. On connaît par ailleurs beaucoup de solutions de l'équation $u^6 + v^6 + w^6 = z^6 + t^6 + 1$ (A. Bremner, 1980), par une méthode inspirée de celle qui est mentionnée pour l'équation $x^4 + y^4 = z^4 + 1$.

5. Équations diophantiennes exponentielles

On appelle ainsi les équations du type :

$$f(x_1^{m_1}, \ldots, x_n^{m_n}) = 0,$$

où f est un polynôme en n variables à coefficients entiers, à résoudre en entiers positifs $(x_1, \ldots, x_n, m_1, \ldots, m_n)$.

Comme équations classiques de ce type, résolubles par des factorisations en nombres entiers, citons :

$$x^2 + 1 = y^n$$

en (x, y, n), qui n'a pas de solution avec $n > 1$ et $y > 1$ (V. A. Lebesgues, 1850), et :

$$x^2 - 1 = y^n,$$

qui n'a pas de solution avec $n > 1$ et $y > 3$ (Chao Ko, 1964).

Pour établir une conjecture de Ramanujan, l'équation :

$$x^2 + 7 = 2^n$$

n'a pour solutions que n = 3, 4, 5, 7. 15, Nage11 (1960) eut recours à la théorie algébrique des nombres (calculs dans le corps Q $(\sqrt{-7})$.

Les résultats d'A. Baker sur les formes linéaires de logarithmes ont permis de réaliser d'importants progrès (cf. nombres TRANSCENDANTS). Il s'agit en fait de méthodes d'approximation.

Ainsi, pour $f(x)$ un polynôme à coefficients entiers avec au moins deux zéros, on sait (A. Schinzel, R. Tijdeman, 1976) qu'il n'y a qu'un nombre fini d'entiers m pour lesquels l'équation :

$$y^m = f(x)$$

a des solutions avec y > 1.

En 1976, R. Tijdeman, utilisant ces méthodes, a montré que l'équation de Catalan :

$$x^m - y^n = 1$$

(en x, y, m, n entiers naturels) n'a qu'un nombre fini de solutions, et ce, par une méthode effective (x^m < exp exp exp 250, Langevin, 1976). La conjecture de Catalan est ainsi ramenée à un nombre fini (mais grand !) de calculs.

MARCEL DAVID
et JEAN-LOUIS COLLIOT-THÉLÈNE

Bibliographie

Z. I. BOREVITH & I. R. CHAFAREVITCH. *Théorie des nombres*, trad. J.-L. Verley, Gauthier-Villars, 1967 / J. W. S. CASSELS, *Rational Quadratic Forms*, Academic Press, San Diego (Calif.), 1979 / A. FAISAN-~, *L 'Équation diophantienne du second degré*, Hermann, 1991 / R. GUY & H. CROFT, *Unsolved Problems in Number Theory*, Springer Verlag, New York-Berlin, 1981 / G. H. HARDY & E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Oxford Univ.. New York, 5' éd. 1979 / K. IRELAND & M. ROSEN. *A Classical Introduction to Modern Number Theory*, Springer-Verlag, New York. 1990 / Yu. I. MANIN. *Cubic Forms : Algebra, Geometry, Arithmetic*, Elsevier Science Publ., New York, 1986 / J. P. MICHON, *Équations diophantiennes*, vol. 1, E.S.T., 1989 / P. RIBENBOIM, 13 *Lectures on Fermat's Last Theorem*, Springer Verlag, New York-Berlin, 1979.

DISTRIBUTIONS

Il est arrivé à plusieurs reprises que certaines exigences de la physique. par exemple, aient conduit les utilisateurs des mathématiques à des « calculs » non rigoureusement justifiables au moyen des concepts mathématiques existants, mais qui traduisaient avec succès la réalité expérimentale. C'est ainsi que l'ingénieur Heaviside introduisit dans l'étude des réseaux électriques (en 1894) les règles de son calcul *symbolique*, qui ne fut justifié mathématiquement que postérieurement. L'étude des équations aux dérivées partielles conduisait aussi naturellement à des extensions des matériaux mathématiques traditionnels ; ainsi, il est normal de considérer que les deux équations :

$$\partial^2 u/\partial x \partial y = 0 \quad \text{et} \quad \partial^2 u/\partial y \partial x = 0$$

sont équivalentes, et pourtant la première est satisfaite par toute fonction $u(x)$ de x seul, alors que l'expression $\partial^2 u/\partial y\ \partial x$ n'a de sens que si $u(x)$ est dérivable en x. Des considérations de ce type, ainsi que l'étude du problème de Dirichlet (trouver une fonction harmonique dans un ouvert de $\mathbf{R}^n$ connaissant ses valeurs sur la frontière) avec les méthodes de l'espace de Hilbert, ont conduit les mathématiciens à généraliser les solutions acceptables d'une telle équation en introduisant la notion de *solution faible*. Le mathématicien soviétique Sobolev a construit, en 1934, des classes de fonctions généralisées qui justifiaient de manière rigoureuse ce genre de considération.

La théorie des transformations de Fourier et de Laplace exigeait aussi des généralisations des fonctions. En 1926, Dirac introduisait en physique mathématique sa

célèbre « fonction » δ_0, nulle en dehors de l'origine et d'intégrale égale à 1, qui représentait une impulsion unité à l'instant $t = 0$, donc d'effet nul pour $t \neq 0$. Puisque δ_0 n'est pas une fonction au sens usuel (car une fonction nulle pour $t \neq 0$ est d'intégrale nulle), sa justification mathématique correcte conduisait à une extension de la notion de fonction ; remarquons que, dans ce cas précis, la théorie de la mesure permettait déjà de considérer δ_0 comme une mesure de masse 1 concentrée à l'origine, c'est-à-dire comme un être mathématique bien défini.

Cette extension a été présentée sous sa forme actuelle par le mathématicien français L. Schwartz, dans le cadre des espaces vectoriels topologiques ; parmi ses nombreuses applications, citons : les équations aux dérivées partielles linéaires, la représentation des groupes de Lie, les processus stochastiques, les variétés différentiables. la physique mathématique, la physique expérimentale (« déconvolution » et identification de systèmes).

La construction des distributions due à L. Schwartz admet de nombreuses variantes conduisant à des classes de fonctions généralisées ayant chacune un domaine privilégié d'applications : fonctions généralisées de divers types introduites par les mathématiciens soviétiques Guelfand et Šilov dans l'étude des équations aux dérivées partielles ; hyperfonctions de Sato-Martineau, très utiles dans l'étude des fonctions de plusieurs variables complexes et les problèmes aux limites ; fonctions généralisées de Beurling-Björk ; etc.

L'exposé qui suit suppose seulement connue la notion d'espace vectoriel (cf. ALGÈBRE ou algèbre LINÉAIRE ET MULTILINÉAIRE) et la notion de suite convergente de nombres complexes (cf. nombres COMPLEXES).



1. Espaces avec notion de suite convergente

Les conditions de continuité qui interviennent dans la définition des distributions peuvent s'exprimer élémentairement en utilisant seulement la notion de suite convergente, sans qu'il soit nécessaire de préciser complètement la topologie des espaces considérés. On se propose ici de montrer comment on peut définir a priori et de manière purement formelle une telle notion sur un espace vectoriel. Les espaces vectoriels sont sur le corps R des nombres réels ou le corps C des nombres complexes.

Définition

Soit E un espace vectoriel. On dit qu'on a défini dans E une *notion de suite convergente* si on s'est donné un sous-ensemble $\mathcal{E}$ de l'ensemble de toutes les suites d'éléments de E et une application de $\mathcal{E}$ dans E qui à toute suite (x_n) de $\mathcal{E}$ fait correspondre un élément $x \in$ E, ce qu'on écrira (de manière purement formelle) : $(x_n) \rightarrow x$ dans E, et ce qu'on lira : La suite (x_n) converge vers x ; les éléments de $\mathcal{E}$ s'appellent *suites convergentes*. On impose aux données précédentes les conditions suivantes :

(*a*) Pour tout élément $x \in$ E, la suite constante $(x, x, \ldots, x, \ldots)$ est convergente et converge vers x ;

(*b*) Si la suite (x_n) est convergente et converge vers x, alors, pour tout nombre A du corps de base R ou C, la suite $(\text{A}\, x_n)$ converge vers $\text{A}\, x$;

(*c*) Si (x_n) et (y_n) sont deux suites convergentes qui convergent respectivement vers x et y, alors la suite $(x_n + y_n)$ converge vers $x + y$.

(*d*) Si (x_n) converge vers x, toute sous-suite de (x_n) converge aussi vers x.

Les conditions ci-dessus sont les propriétés des suites convergentes (au sens usuel) de nombres réels ou complexes. Comme toujours dans l'approche formelle d'une notion, on retrouve donc, sous forme d'axiomes, des propriétés vérifiées dans les situations concrètes qu'il s'agit de généraliser. Si une suite (x_n) converge vers x, on dit aussi que (x_n) a pour limite x et on écrit :

$$\lim_{n \to \infty} x_n = x$$

Remarquons que, pour connaître $\mathcal{E}$, il suffit de connaître le sous-ensemble $\mathcal{E}_0$ de $\mathcal{E}$ formé des suites (x_n) qui convergent vers 0 (d'après les axiomes, c'est d'ailleurs un espace vectoriel pour les opérations usuelles sur les suites). En effet, dire que $(x_n) \rightarrow x$ équivaut, d'après les axiomes, à dire que la suite $(x_n - x)$ tend vers 0, ce qui met en évidence que la translation de vecteur x, qui à x_n fait correspondre $x_n + x$, est une bijection de $\mathcal{E}_0$ sur l'ensemble $\mathcal{E}_x$ des suites qui convergent vers x. En abrégé, on dira qu'un espace vectoriel E est un e.v.s. si on a défini dans E une notion de suite convergente.

Un exemple fondamental

11 est clair que, si E est l'espace euclidien usuel de la géométrie dans l'espace. l'ensemble des suites convergentes au sens usuel satisfait aux conditions précédentes ; plus généralement, si E est un espace vectoriel normé, muni d'une norme $\|.\|$, on peut définir directement, à partir de la norme, les suites (x_n) qui convergent vers x par la propriété suivante : La suite de nombres réels positifs $\|x - x_n\|$ tend vers 0 pour $n \rightarrow \infty$; il est clair que les conditions (a) à (*d*) sont alors satisfaites. L'exemple suivant est essentiel dans la définition des distributions ; on remarquera qu'on définit ici les suites convergentes sans l'intermédiaire d'une topologie.

Soit Ω un sous-ensemble ouvert de $\mathbf{R}^n$ (c'est-à-dire que pour tout point de Ω il existe une boule de rayon > 0 contenue dans Ω). Toutes les fonctions considérées sont supposées à valeurs complexes. Si φ est une telle fonction définie et continue dans Ω, on appelle *support* de φ le plus petit ensemble fermé en dehors duquel φ est nulle ; $\mathfrak{D}(\Omega)$ *désignera l'ensemble des fonctions définies dans* Ω, *admettant des dérivées partielles de tous ordres, et à support compact* (c'est-à-dire borné et fermé dans $\mathbf{R}^n$).

Pour désigner les dérivées partielles d'ordre quelconque, on utilise la convention des multi-indices (cf. CALCUL INFINITÉSIMAL - Calcul à plusieurs variables). Par définition. un multi-indice est un système de n nombres entiers positifs ou nuls :

$$\alpha = (\alpha_1, \alpha_2, ..., \alpha_n) ;$$

on écrit alors $\partial^\alpha/\partial x^\alpha$ ou $(\partial/\partial x)^\alpha$ pour désigner l'opérateur de dérivation partielle

$$(\partial^{\alpha_1 + \alpha_2 + \dots + \alpha_n})/\partial x_1{}^{\alpha_1} \partial x_2{}^{\alpha_2} \dots \partial x_n{}^{\alpha_n} = (\partial/\partial x_1)^{\alpha_1}(\partial/\partial x_2)^{\alpha_2} \dots (\partial/\partial x_n)^{\alpha_n} ;$$

cette écriture permet d'avoir, dans le cas de n variables. une écriture analogue au cas d'une variable.

11 est clair que B(Q) est un espace vectoriel ; munissons-le d'une structure d'e.v.s. en définissant les suites convergentes. Soit (φ_p) une suite d'éléments de $\mathfrak{D}(\Omega)$; on dira que la suite (φ_p) est convergente et converge vers une fonction $\varphi \in \mathfrak{D}(\Omega)$ si les deux conditions suivantes sont réalisées :

(*a'*) Toutes les fonctions φ_p, ainsi que la fonction φ, sont nulles en dehors d'un même compact K de Ω ;

(h') Pour tout multi-indice α, la suite des dérivées partielles $(\partial^\alpha \varphi_p / \partial x^\alpha)$ converge uniformément sur K vers la dérivée partielle correspondante de φ :

$$\frac{\partial^\alpha \varphi_p}{\partial x^\alpha} \rightarrow \frac{\partial^\alpha \varphi}{\partial x^\alpha},$$

pour $p \rightarrow \infty$. Il est clair que les conditions (a) à (d) sont satisfaites.

Morphismes

On va maintenant définir les morphismes des e.v.s., c'est-a-dire les applications d'un tel e.v.s. dans un autre qui respectent les deux notions définissant la structure d'un e.v.s. : la structure vectorielle et les « suites convergentes ».

Soit E et F. deux e.v.s. Un morphisme u de E dans F est, par définition, une application linéaire de E dans F (c'est-à-dire telle que $u(\lambda x + \mu y) = \lambda u(x) + \mu u(y)$, pour $x, y \in$ E et λ, μ dans le corps de base R ou C) qui transforme toute suite convergente de E en une suite convergente de F : si $(x_n) \rightarrow x$ dans E, alors $(u(x_n)) \rightarrow u(x)$ dans F ; on dit aussi que u est une application linéaire séquentiellement continue (ou « continue pour les suites »). Pour qu'une application linéaire de E dans F soit un morphisme, il faut et il suffit qu'elle transforme toute suite convergente vers 0 dans E en une suite convergente vers 0 dans F.

Voici deux exemples de morphismes de l'e.v.s. $\mathfrak{D}(\Omega)$ dans lui-même. Si on désigne par $\partial/\partial x_i$ l'opérateur de dérivation partielle par rapport à la i-ième coordonnée dans $\mathbf{R}^n$, l'application $\varphi \mapsto \partial \varphi/\partial x_i$ est un morphisme de $\mathfrak{D}(\Omega)$. De même, si f est une fonction admettant dans Ω des dérivées partielles de tous ordres, l'opération de multiplication par f, qui s'écrit $\varphi \mapsto f\varphi$, est un morphisme de $\mathfrak{D}(\Omega)$ dans $\mathfrak{D}(\Omega)$.

Si E et F sont deux e.v.s., on peut munir l'espace vectoriel $\mathfrak{L}$ (E, F) des morphismes de E dans F d'une notion de suite convergente en disant qu'une suite u_n de morphismes de E dans F converge vers $u \in \mathfrak{L}(E, F)$ si $(u_n(x)) \rightarrow u(x)$ dans F pour tout élément $x \in$ E. En particulier, cette définition permet de munir d'une structure d'e.v.s. l'espace vectoriel E' des applications linéaires de E dans son corps de base qui sont continues pour les suites.

On retrouve dans le cadre des e.v.s. l'importante notion de transposée d'une application linéaire (cf. algèbres LINÉAIRE ET MULTILINÉAIRE). Soit, en effet, E et F deux e.v.s., et u un morphisme de E dans F ; désignons par E' et F', comme ci-dessus, les e.v.s. des applications linéaires respectivement de E et F dans le corps de base (formes linéaires sur E et F), qui sont séquentiellement continues. Pour toute forme linéaire $f \in$ F', la forme $g = f \circ u$ est une forme linéaire séquentiellement continue sur E, donc est un élément de E', et on vérifie facilement que l'application linéaire ${}^t u$: F' $\rightarrow$ E', qui à $f \in$ F' fait correspondre $f \circ u \in$ E', est séquentiellement continue ; on définit ainsi le *morphisme* ${}^t u$ *transposé* de u.

2. Définition des distributions

Il est clair que, pour généraliser la notion de fonction, il faut abandonner certaines propriétés usuelles des fonctions (par exemple le fait qu'une fonction prend une valeur déterminée en chaque point) pour ne conserver que certaines propriétés. L. Schwartz utilise comme notion essentielle la propriété d'opérer linéairement sur des classes de fonctions très régulières.

Méthode générale de construction

La méthode générale pour définir un espace de fonctions généralisées sur un ouvert Ω de $\mathbf{R}^n$ est la suivante. On prend tout d'abord un e.v.s. $\mathfrak{C}$ de fonctions « suffisamment régulières » dans Ω ; par définition, l'espace des fonctions généralisées sur Ω est alors l'espace $\mathfrak{C}'$ des formes linéaires séquentiellement continues sur $\mathfrak{C}$. Pour justifier la terminologie de « fonctions généralisées », il faut identifier les fonctions « régulières » sur Ω à des éléments de $\mathfrak{C}'$; pour cela, on identifie une telle fonction f à la forme linéaire sur $\mathfrak{C}$:

$$\varphi \mapsto (f, \varphi) = \int_\Omega f(x)\varphi(x)\,dx, \quad \varphi \in \mathfrak{C},$$

où $x = (x_1, ..., x_n)$ et $dx = dx_1\,dx_2 \ldots dx_n$. Nous allons préciser ces indications générales en construisant en détail les distributions proprement dites.

Définition des distributions

On prend ici pour espace $\mathfrak{C}$ l'e.v.s. $\mathfrak{D}(\Omega)$, défini ci-dessus, des fonctions indéfiniment différentiables et à support compact dans Ω. L'espace $\mathfrak{D}'(\Omega)$ des *distributions* dans Ω est alors par définition l'ensemble des formes linéaires T séquentiellement continues sur $\mathfrak{D}$; on note indifféremment :

$$T(\varphi) = (T, \varphi) = \langle T, \varphi \rangle = \int_\Omega T(x)\varphi(x)\,dx,$$

la valeur de la distribution T (forme linéaire sur $\mathfrak{D}(\Omega)$) sur la fonction $\varphi \in \mathfrak{D}(\Omega)$. Remarquons que la dernière écriture est abusive ; elle est utilisée car elle rappelle que, si T est une fonction, l'expression de $T(\varphi)$ est donnée par une intégrale. Précisons ce point.

Si f est une fonction intégrable (au sens de la théorie de Lebesgue) sur tout compact de Ω (on dit alors que f est localement intégrable), on peut l'identifier à la distribution :

$$\varphi \mapsto \int f(x)\varphi(x)\,dx, \quad \varphi \in \mathfrak{D}(\Omega);$$

les distributions définies par deux fonctions f et g localement intégrables dans Ω coïncident si et seulement si f et g « coïncident » au sens de la théorie de Lebesgue, c'est-à-dire sont *égales presque partout* ; ainsi, les distributions ne généralisent pas à proprement parler les fonctions, mais les classes de fonctions égales presque partout : si on modifie une fonction en changeant sa valeur en un point, par exemple, elle définit toujours la même distribution ; on a bien abandonné la propriété des fonctions d'être définies par leur valeur en *tout* point.

La notion d'e.v.s. permet de définir la notion de *suites convergentes de distributions* : si T_1, T_2, ..., T_n, est une suite de distributions, on dit, en accord avec la définition de l'e.v.s. $\mathfrak{D}'(\Omega)$, que cette suite tend vers une distribution T si :

$$T(\varphi) = \lim_{n\to\infty} T_n(\varphi),$$

pour toute fonction $\varphi \in \mathfrak{D}(\Omega)$. On peut montrer que, si pour toute fonction $\varphi \in \mathfrak{D}(\Omega)$ la suite de nombres complexes $T_n(\varphi)$ tend vers une limite, l'application définie dans $\mathfrak{D}(\Omega)$ qui à $\varphi \in \mathfrak{D}(\Omega)$ fait correspondre cette limite est une distribution, limite de la suite des distributions T_n. Cette propriété est généralement très facile à vérifier et permet de définir de nombreuses distributions nouvelles à partir de distributions déjà connues.

La notion de suite convergente de distributions permet en particulier de définir les *séries convergentes* de distributions. Si (T_n) est une suite de distributions, on dit que la série de terme général T_n est

convergente (au sens des distributions) de somme T si la suite des sommes partielles est convergente au sens indiqué ci-dessus ; on écrit alors $\Sigma T_n = T$.

Exemples

a) Soit a un point de $\mathbf{R}^n$; l'application qui à toute fonction $\varphi \in \mathfrak{D}(\mathbf{R}^n)$ fait correspondre la valeur $\varphi(a)$ de la fonction φ en a est une distribution appelée *distribution de Dirac* et notée 6,. Ainsi, avec l'abus d'écriture signalé ci-dessus, on a :

$$(\delta_a, \varphi) = \int \varphi(x)\,\delta_a(x)\,dx = \varphi(a)\,;$$

cette distribution permet de donner une définition mathématique rigoureuse de la « fonction » de Dirac mentionnée plus haut.

b) On vérifie facilement que la fonction *a* définie sur R par

$$a(x) = \begin{cases} \exp\,[-1/(1-x^2)] & \text{si } |x| < 1, \\ 0 & \text{si } |x| \geqslant 1, \end{cases}$$

est une fonction indéfiniment dérivable, nulle par définition en dehors de l'intervalle [- 1, 1]. Si on pose :

$$k = \int_{-1}^{+1} a(x)\,dx,$$

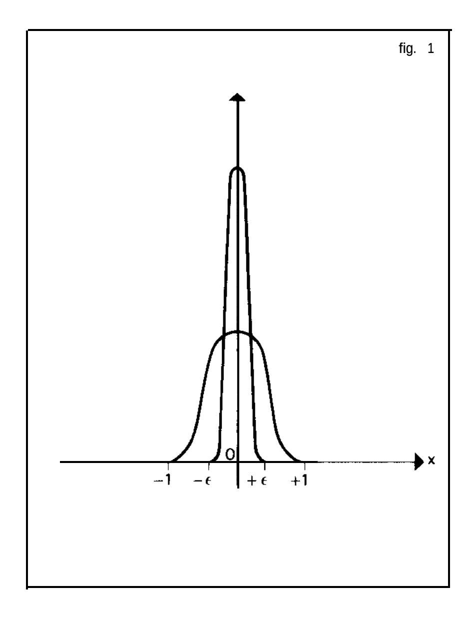
la fonction $\rho_1(x) = a(x)/k$ appartient donc à $\mathfrak{D}(\mathbf{R})$ et est d'intégrale égale à 1. Plus généralement, la fonction :

$$\rho_\varepsilon(x) = (1/\varepsilon)\,\rho(x/\varepsilon)$$

appartient aussi à $\mathfrak{D}(\mathbf{R})$ et est d'intégrale égale à 1 ; d'autre part, ρ_ε est nulle en dehors de l'intervalle $[-\varepsilon, \varepsilon]$ (fig. 1). Comme

$$\int \rho_\varepsilon(x)\,\varphi(x)\,dx \to \varphi(0), \text{ si } \varepsilon \to 0,$$

pour toute fonction continue φ, il en résulte que, pour toute suite E,, tendant vers 0, la suite des distributions $\rho_{\varepsilon n}$ (ou, pour être précis, la suite des distributions T_n définies par ces fonctions) tend vers la distribution de Dirac au point 0.



c) Dans le plan, identifié à $\mathbf{R}^2$, de la variable complexe z, on montre que la fonction $f(x + iy) = 1/\pi z$ est intégrable sur tout compact ; par suite, cette fonction définit une distribution, notée $1/\pi z$, qui, à toute fonction $\varphi \in \mathfrak{D}(\mathbf{R}^2)$, associe :

$$(1/\pi z, \varphi) = \iint_{\mathbf{R}^2} \frac{\varphi(x, y)}{\pi(x + iy)}\,dx\,dy.$$

d) Soit maintenant $\mathbf{R}^3$ l'espace de la géométrie élémentaire dans l'espace. La fonction $x \mapsto 1/(4\pi\|x\|)$ est localement intégrable dans $\mathbf{R}^3$ et définit donc une distribution, notée $1/(4\pi\|x\|)$.

e) On peut montrer que, pour toute fonction $\varphi \in$ 9(R), et pour toute suite E,, tendant vers 0, la suite des nombres réels :

$$\int_{|x| > \varepsilon_n} \frac{\varphi(x)}{x}\,dx$$

tend vers une limite, indépendante de la suite (ε_n) choisie, et que l'application qui à φ fait correspondre la limite correspondante est une distribution, notée v.p.($1/x$), valeur principale de Cauchy ; ainsi :

$$(\text{v.p.}(1/x), \varphi) = \lim_{\varepsilon \to 0} \int_{|x| > \varepsilon} \frac{\varphi(x)}{x} dx.$$

Dans l'étude des équations aux dérivées partielles, Hadamard a été conduit à généraliser cette notion (distributions « parties finies »).

3. Propriétés des distributions

À partir d'une application linéaire séquentiellement continue L de $\mathfrak{D}$ dans $\mathfrak{D}$ (opérateur dans $\mathfrak{D}$), on peut définir, par transposition, un opérateur linéaire L' dans l'espace $\mathfrak{D}'$ des distributions :

$$(\text{L}'(\text{T}), \varphi) = (\text{T}, \text{L}(\varphi)), \quad \varphi \in \mathfrak{D}.$$

Dérivotion des distributions

Les distributions étant une généralisation de la notion de fonction régulière, essayons d'étendre aux distributions la notion de *dérivation*. Pour cela, analysons tout d'abord les propriétés des opérateurs de dérivation partielle pour les fonctions contmument dérivables dans un ouvert Ω de $\mathbf{R}^n$.

Pour toute fonction $\varphi \in \mathfrak{D}(\Omega)$, on a, si T désigne à la fois une fonction continûment dérivable dans $\mathbf{R}^n$ et la distribution qu'elle définit :

$$\left(\frac{\partial \text{T}}{\partial x_j}, \varphi\right) = \int \frac{\partial \text{T}}{\partial x_j}(x)\varphi(x)\,dx, \qquad x = \left(x_1, x_2, ..., x_n\right);$$

mettant en évidence la variable x_j, on peut écrire (en permutant les variables) x = (x'. x_j), d'où $dx = dx'\,dx_j$, ce qui donne, par intégration par parties,

$$\left(\frac{\partial \text{T}}{\partial x_j}, \varphi\right) = \int dx' \int \frac{\partial \text{T}}{\partial x_j}(x', x_j)\varphi(x', x_j)\,dx_j = -\int dx' \int \text{T}(x)\frac{\partial \varphi}{\partial x_j}(x)\,dx_j = -\left(\text{T}, \frac{\partial \varphi}{\partial x_j}\right).$$

Or, l'opérateur L = $-$ $(\partial/\partial x_j)$ est séquentiellement continu dans $\mathfrak{D}(\Omega)$ (cf. *supra*, *Morphismes)* ; l'opérateur transposé, noté $\partial/\partial x_j$ est donc un opérateur séquentiellement continu de $\mathfrak{D}'(\Omega)$ qui prolonge la dérivation au sens usuel. Pour **toute** fonction φ de $\mathfrak{D}(\Omega)$, laj-ième dérivée partielle de la distribution T donne donc *par déjnition* :

$$\left(\frac{\partial \text{T}}{\partial x_j}, \varphi\right) = -\left(\text{T}, \frac{\partial \varphi}{\partial x_j}\right).$$

Ainsi, par exemple, la distribution dérivée de la distribution de Dirac en α sur la droite est telle que $(\delta'_\alpha, \varphi) = \quad \varphi'(\alpha)$.

support

Si Ω' est un ouvert contenu dans Ω, on a un morphisme *naturel*

$$p : \mathfrak{D}(\Omega') \rightarrow \mathfrak{D}(\Omega),$$

qui à toute fonction $\varphi \in \mathfrak{D}(\Omega')$ associe la fonction $p(\varphi)$ obtenue en prolongeant φ par 0 en dehors de Ω'. L'opérateur trans**posé** de p est l'opérateur p' sur les distributions défini par :

$$(p'(\text{T}), \varphi) = (\text{T}, p(\varphi)), \ \text{T} \in \mathfrak{D}'(\Omega), \ \varphi \in \mathfrak{D}(\Omega');$$

p' (T) est généralement noté $\text{T}_{\Omega'}$, *restriction* de la distribution T à l'ouvert Ω'. On dit que T est *nulle sur* Ω' si $\text{T}_{\Omega'} = 0$.

Si T est une distribution dans un ouvert Ω, on appelle *support* de T le complémentaire dans Ω de la réunion des ouverts Ω' sur lesquels T est nulle. Par exemple, le support de la distribution de Dirac au point a de $\mathbf{R}^n$ est constitué du seul point a.

Produit direct

Soit Ω un ouvert de l'espace $\mathbf{R}^n$ de la variable x, et Ω' un ouvert de l'espace $\mathbf{R}^p$ de la variable y, respectivement. Si T et U sont des fonctions continues dans Ω et Ω' respectivement, la distribution T X U dans $\Omega \times \Omega'$ définie par la fonction T(x) U(y) satisfait à :

$$(T \times U, \varphi) = \iint_{\Omega \times \Omega'} T(x)\, U(y)\, \varphi(x,y)\, dx\, dy$$
$$(1) \qquad = \int_{\Omega} T(x)\, dx \int_{\Omega'} U(y)\, \varphi(x,y)\, dy$$
$$(2) \qquad = \int_{\Omega'} U(y)\, dy \int_{\Omega} T(x)\, \varphi(x,y)\, dx,$$

pour toute fonction $\varphi \in \mathfrak{D}(\Omega \times \Omega')$.

On peut montrer que lorsque T et U sont des distributions, si on considère les intégrales comme des accouplements distributions-fonctions (abus de notation signalé plus haut), alors les expressions (1) et (2) ont un sens, sont égales, et on peut montrer que l'application qui à $\varphi \in \mathfrak{D}(\Omega \times \Omega')$ fait correspondre (1) ou (2) est une distribution. Cette distribution, notée T X U, est appelée produit *direct* de T et U.

Convolution

Soit T et U deux fonctions continues et intégrables dans $\mathbf{R}^n$; on appelle *produit de convolution* de T et U la fonction définie par la formule intégrale

$$T * U(x) = \int_{\mathbf{R}^n} T(x-y)\, U(y)\, dy.$$

La fonction T $*$ U ainsi définie est telle que, pour toute fonction φ de $\mathfrak{D}(\mathbf{R}^n)$,

$$(3) \quad (T * U, \varphi) = \int dx\, \varphi(x) \int T(x-y) U(y)\, dy$$
$$= \iint_{\mathbf{R}^n \times \mathbf{R}^n} T(x)\, U(y)\, \varphi(x+y)\, dx\, dy.$$

On montre que, dans certains cas, on peut donner un sens à (3) pour des distributions T et U, et définir ainsi une distribution, notée T $*$ U. Cela est possible, par exemple, si l'une des distributions est à support compact. Sur R, on peut également définir le produit de convolution de deux distributions S et T qui s'annulent pour $t < 0$.

L'intérêt pratique de la convolution est que de nombreuses opérations usuelles sont des convolutions. Par exemple, pour toute distribution T sur $\mathbf{R}^n$ et pour tout multi-indice α :

$$\left(\frac{\partial}{\partial x}\right)^{\alpha} T = \left(\frac{\partial}{\partial x}\right)^{\alpha} \delta_0 * T,$$

où δ_0 est la distribution de Dirac à l'origine des coordonnées. On voit l'intérêt de l'étude des *équations de convolution*, du type A $*$ X = B, où A et B sont des distributions connues ; la résolution d'une telle équation est parfois appelée « déconvolution ». Si A est une distribution à support compact, on appelle *solution élémentaire* de A toute distribution E telle que E $*$ A = A $*$ E = δ_0 ; une solution élémentaire d'un opérateur différentiel

$$P\left(\frac{\partial}{\partial x}\right) = \sum_{|\alpha| \leqslant m} a_\alpha \left(\frac{\partial}{\partial x}\right)^{\alpha}$$

est alors par définition une solution élémentaire de la distribution A = $P(\partial/\partial x)\delta_0$.

4. Séries et intégrales de Fourier

La plupart des grandes théories de l'analyse classique s'étendent aux distributions ; nous nous limiterons ici à des indications rapides sur la théorie de Fourier (cf. analyse HARMONIQUE) en renvoyant à l'article calcul SYMBOLIQUE pour la transformation de Laplace.

Transformation de Fourier dans l'espace S

Le domaine naturel de la transformation de Fourier élémentaire est l'espace S des fonctions indéfiniment dérivables à décroissance rapide ainsi que toutes leurs dérivées : c'est l'espace des applications φ indéfiniment dérivables de $\mathbf{R}^n$ dans le corps C des nombres complexes telles que :

$$\|\varphi_n\|_{k,l} = \sup_{x \in \mathbf{R}^n} (1 + \|x\|^2)^l \left| \left(\frac{\partial}{\partial x} \right)^k \varphi(x) \right| < \infty,$$

pour tout entier l et tout multi-indice k. Cet espace $\mathcal{S}(\mathbf{R}^n)$ est une e.v.s. pour la notion suivante de suites convergentes : par définition, une suite (φ_n) de fonctions de S tend vers 0 si :

$$\|\varphi_n\|_{k,l} \to 0, \quad \text{pour } n \to \infty,$$

quels que soient l'entier /et le multi-indice k.

Pour $\varphi \in \mathcal{S}(\mathbf{R}^n)$, on appelle *transformée de Fourier* de $\mathcal{S}$, la fonction :

$$\hat{\varphi}(u), u = (u_1, u_2, \dots u_n) \in \mathbf{R}^n,$$

définie par la formule intégrale :

$$\hat{\varphi}(u) = \int_{\mathbf{R}^n} \varphi(x)\, e^{-i<x,u>}\, dx,$$

ou $< x, u > = u_1 x_1 + u_2 x_2 + \quad + u_n x_n$ désigne le produit scalaire dans R". Cette définition de la transformation de Fourier n'est pas universelle ; certains auteurs préfèrent prendre par exemple :

$$\hat{\varphi}(u) = \int \varphi(x)\, e^{-2i\pi<x,u>}\, dx,$$

ou $$\hat{\varphi}(u) = \int \varphi(x) e^{i<x,u>}\, dx, \ \dots$$

On passe d'une définition à l'autre par un simple changement de variable dans l'intégrale définissant $\hat{\varphi}$. Nous écrirons **tout** ce qui suit avec la convention indiquée, adoptée par de nombreux mathématiciens qui étudient les équations aux dérivées partielles. Dans l'article analyse HARMONIQUE, au contraire, on préférera adopter la seconde formule.

L'application $\mathcal{F}$ qui à φ fait correspondre sa transformée de Fourier $\hat{\varphi}$ possède les principales propriétés suivantes :

u) la fonction φ est dans S et

$$\mathcal{F} : \mathcal{S} \to \mathcal{S}$$

est un isomorphisme de l'e.v.s. S sur lui-même ;

b) pour tout couple de fonctions φ et ψ de S, on a les formules :

$$\int \varphi(x)\, \hat{\psi}(x)\, dx = \int \hat{\varphi}(x)\, \psi(x)\, dx$$

(relation de Parseval),

$$\int <p(x)\ \psi(x)\, dx = (2\pi)^{-n} \int \hat{\varphi}(u)\, \hat{\psi}(u)\, du$$

(relation de Plancherel) ;

c) l'isomorphisme réciproque de l'isomorphisme $\mathcal{F}$ est défini par :

$$\varphi(x) = (2\pi)^{-n} \int \hat{\varphi}(u)\, e^{i<x,u>}\, du$$

(formule d'inversion de Fourier).

Transformation de Fourier dans S'

On appelle *distribution tempérée* dans $\mathbf{R}^n$ toute forme linéaire séquentiellement continue sur S ; remarquons que, puisque l'application identique de D(R") dans $\mathcal{S}(\mathbf{R}^n)$ est un morphisme et que toute φ de S est limite d'une suite d'éléments de $\mathcal{D}$, la transposée de cette application identique est une injection. Cette injection fait apparaître l'espace S' des distributions tempérées comme un sous-espace vectoriel de l'espace vectoriel $\mathcal{D}'$ des distributions.

Par définition, on appelle alors *transformée de Fourier d'une distribution tem-*

pérée T la distribution tempérée $\hat{T}$ définie par :

$$(\hat{T}, \varphi) = (T, \hat{\varphi}), \quad \varphi \in \mathcal{S}.$$

Considérons par exemple la distribution sur R définie par la fonction x ; sa transformée de Fourier $\hat{x}$ est telle que :

$$(\hat{x}, \varphi) = (x, \hat{\varphi}) = \int u \hat{\varphi}(u)\, du\ ;$$

un calcul facile montre que cette dernière intégrale vaut :

$$-2i\pi\varphi'(0) = 2\, i\pi(\delta'_0, \varphi),$$

où δ'_0 est la dérivée (au sens de distributions) de la distribution de Dirac en 0. Ainsi $\hat{x} = 2\, i\, \pi\delta'_0$. Par des raisonnements analogues, on pourrait montrer que la distribution :

$$T = \sum_{n=-\infty}^{\infty} \delta_n$$

a pour transformée de Fourier :

$$\hat{T} = 2\pi \sum_{n=-\infty}^{\infty} \delta_{2n\pi},$$

ce qui est équivalent a la relation de Poisson,

$$\sum_{n=-\infty}^{+\infty} \varphi(n) = 2\pi \sum_{n=-\infty}^{+\infty} \hat{\varphi}(2\pi n),$$

valable pour toute fonction $\varphi \in S(R)$.

Coefficients de Fourier d'une distribution périodique

Comme d'habitude, on identifiera le tore $T = \mathbf{R}/2\pi\mathbf{Z}$ (quotient de R par la relation d'équivalence $x \sim y$ six y est un multiple entier de 2π) à la circonférence unité du plan complexe, tout point de cette circonférence étant caractérisé par son affixe $e^{i\theta}$, ou, ce qui revient au même, par son abscisse curviligne θ. Cette identification établit une correspondance naturelle entre les fonctions définies sur le tore T et les fonctions périodiques de période 2π définies sur R. Par définition, on dit qu'une fonction définie sur T est indéfiniment dérivable s'il en est ainsi pour la fonction périodique associée ; l'espace $\mathcal{D}(\mathbf{T})$ des fonctions indéfiniment dérivables sur T est alors muni d'une structure d'e.v.s. en convenant qu'une suite (φ_n) tend vers 0 dans 9(T) pour $n \to \infty$, si $(d/d\theta)^k \varphi_n$ tend uniformément vers 0 sur T, $n \to \infty$, pour tout entier k.

Par définition, une *distribution* sur T est alors une forme linéaire séquentiellement continue sur $\mathcal{D}(\mathbf{T})$; il est facile de voir que l'espace 9'(T) de ces distribution est en correspondance bijective avec l'ensemble des *distributions périodiques* sur R, c'est-a-dire les distributions T telles que :

$$(T, q(x)) = (T, \varphi(x - 2\pi)), \quad \varphi \in \mathcal{D}(\mathbf{R}).$$

En procédant comme ci-dessus, on peut définir des opérations dans $\mathcal{D}'(\mathbf{T})$: dérivation, produit direct, convolution, etc. Si fest une fonction intégrable définie sur T, elle définit la distribution :

$$\int_0^{2\pi} f(\theta)\, \varphi(\theta) \frac{d\theta}{2\pi}.$$

Par définition, on appelle *transformée de Fourier* d'une fonction $\varphi \in \mathcal{D}(\mathbf{T})$ la suite doublement infinie $\hat{\varphi} = (\hat{\varphi}_k)_{-\infty<k<\infty}$ définie par :

$$\hat{\varphi}_k = \int_0^{2\pi} \varphi(\theta) e^{-ik\theta} \frac{d\theta}{2\pi}, \quad k \in \mathbf{Z}.$$

Pour développer une théorie similaire à celle de la transformation de Fourier vue ci-dessus, on est conduit à introduire l'espace vectoriel s des suites doublement infinies à décroissance rapide, c'est-à-dire les suites $a = (a_k)_{k\in Z}$ telles que :

$$\|a_k\|_l = \sup_k (1 + |k|)^l |a_k| < \infty,$$

pour tout entier l ; cet espace est un e.v.s. si on convient qu'une suite $(a^{(n)})$ d'éléments de s tend vers 0 si pour tout entier l la suite $\| a^{(n)} \|_l$ tend vers 0 pour $n \to \infty$.

On montre alors que l'application $\mathcal{F}$ définie par $\mathcal{F}(\varphi) = \hat{\varphi}$ est un isomorphisme de $\mathfrak{D}(\mathbf{T})$ sur l'espace s, l'application inverse faisant correspondre à la suite a = (a,) la somme de la série de Fourier :

$$\varphi(\theta) = \sum_{k=-\infty}^{+\infty} a_k e^{ik\theta}.$$

La relation de Plancherel prend ici la forme :

$$\int_0^1 \varphi(\theta)\psi(\theta)\frac{d\theta}{2\pi} = \sum_{k=-\infty}^{+\infty} \hat{\varphi}_k \hat{\psi}_{-k},$$

$$\varphi, \psi \in \mathfrak{D}(\mathbf{T}) ;$$

elle montre que la transposée de $\mathcal{F}^{-1}$ est un isomorphisme de W(T) surs' qui prolonge $\mathcal{F}$. Si on note encore $\mathcal{F}$ cette application transposée T $\mapsto$ T, on a :

$$(\hat{\mathbf{T}}, \varphi) = \sum_{k=-\infty}^{+\infty} \hat{\mathbf{T}}_k \varphi_{-k} ;$$

en particulier, si on fait $\varphi(\theta) = e^{ik\theta}$, on obtient :

$$\hat{\mathbf{T}}_k = (\mathbf{T}, e^{-ik\theta}).$$

On peut caractériser l'espaces'des transformées de Fourier des distributions tempérées ; c'est l'ensemble des suites T = (T,) *à croissance lente,* c'est-à-dire pour lesquelles il existe un entier l (dépendant de T) tel que T_k/k^l reste borné pour $k \mid \to \infty$.

Une application

Les distributions peuvent servir à étudier le comportement des fonctions analytiques ; voici un exemple simple d'une telle situation.

À une distribution T sur le tore, associons le couple $\mathfrak{T} = (\mathfrak{T}^+, \mathfrak{T}^-)$ des deux fonctions holomorphes ainsi définies :

$$\mathfrak{T}^+(z) = \sum_{k=0}^{\infty} \hat{\mathbf{T}}_k z^k \text{ si } |z| < 1,$$

$$\mathfrak{T}^-(z) = \sum_{k=-\infty}^{-1} \hat{\mathbf{T}}_k z^k \text{ si } |z| > 1.$$

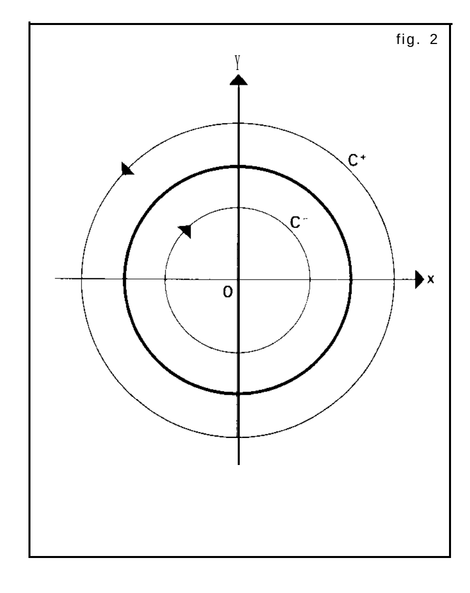
Soit maintenant $\mathrm{P}(\theta)$ un polynôme trigonométrique, c'est-à-dire une combinaison linéaire finie d'exponentielle $e^{ik\theta}$:

$$\mathrm{P}(\theta) = \sum_{|k| \leqslant \mathrm{N}} \hat{\mathrm{P}}_k e^{ik\theta} ;$$

on peut écrire aussi, par abus de notation :

$$\mathrm{P}(z) = \sum \hat{\mathrm{P}}_k z^k.$$

Soit C^+ et C^- deux circonférences de centre 0 et de rayons < 1 et > 1, orientées dans le sens rétrograde et dans le sens trigonométrique respectivement. Développant $\mathfrak{T}$ et P en série, on obtient *(fig. 2) :*



$$\frac{1}{2i\pi}\int_{C+} \mathcal{T}^+(z)\,P(z)\,dz = \sum_{k=0}^{\infty} \hat{T}_k \hat{P}_{-k}$$

$$\frac{1}{2i\pi}\int_{C-} \mathcal{T}^-(z)\,P(z)\,dz = \sum_{k=-\infty}^{-1} \hat{T}_k \hat{P}_{-k}$$

donc :

$$(4)\qquad (T, P) = \frac{1}{2i\pi}\int_{C+\cup C-} \mathcal{T}(z)\,P(z)\,dz.$$

Pour $0 < r < 1$, posons maintenant :

$$T_r^+(\theta) = \mathcal{T}^+(re^{i\theta}) = \sum_{k=0}^{\infty} \hat{T}_k r^k e^{ik\theta}$$

$$T_r^-(\theta) = \mathcal{T}^-\left(\frac{e^{i\theta}}{r}\right) = \sum_{k=-\infty}^{-1} \hat{T}_k r^k e^{ik\theta};$$

ainsi, on a : $T_r^+ - T_r^- = P_r * T$,

où : $$P_r(\theta) = \sum_{k=-\infty}^{+\infty} r^{|k|} e^{ik\theta}$$

est le noyau de Poisson. On peut voir que P_r tend vers δ_0 dans W(T) pour $r \to 1$; il en résulte que l'on a dans W(T) :

$$T = \lim_{r\to 1} (T_r^+ - T_r^-).$$

On interprète le résultat précédent en disant que la distribution T est la différence des *valeurs au bord (le* long de la circonférence unité) des fonctions holomorphes $\mathcal{T}^+$ et $\mathcal{T}^-$. Par exemple. considérons :

$$\mathcal{T}(z) = \frac{1}{1 - z^N},\ \text{N entier positif.}$$

Développant $\mathcal{T}$ en série entière (en z ou en $1/z$), on voit que à $\mathcal{T}$ est associée une distribution T_N dont les coefficients de Fourier sont :

$$(\hat{T}_N)_k = \begin{cases} 0, \text{ si } k \text{ n'est pas un multiple de N,} \\ 1, \text{ si } k \text{ est un mutiple de N ;} \end{cases}$$

la formule (4) ci-dessus montre alors que pour tout polynôme trigonométrique P :

$$(T_N, P) = \sum_{j=0}^{n-1} P\left(z_N^j\right), \quad \text{avec } z_N^j = e^{2\pi i j/N}.$$

Si δ_N^j désigne la mesure de Dirac sur T concentrée au point z_N^j, on a donc :

$$(\hat{T}_N, P) = \sum_{j=0}^{n-1} \left(\delta_N^j, P\right);$$

comme toute fonction de B(T) est approchable uniformément par une suite de polynômes trigonométriques, on en déduit que T_N est la somme de la série :

$$T_N = \sum_{j=0}^{n-1} \delta_N^j.$$

De manière générale, toute suite doublement infinie $(\hat{a}_k)_{k\in\mathbf{Z}}$, à croissance lente (cf. *supra*, *Coefficients de Fourier d'une distribution périodique*), est la suite des coefficients de Fourier d'une distribution A sur T, ce qu'on écrit *symboliquement* :

$$A(\theta) \sim \sum_{k=-\infty}^{+\infty} \hat{a}_k e^{ik\theta};$$

de plus, A est la limite dans W(T) des distributions définies par les polynômes trigonométriques :

$$\sum_{|k|\leqslant N} \hat{a}_k e^{ik\theta}, \quad N \to \infty.$$

Autrement dit. la série de Fourier d'une distribution converge vers cette distribution dans B'(T) ; mais bien entendu cela n'entraîne pas que les sommes partielles convergent pour tout θ.

PAUL KRÉE

Bibliographie

J. BARROS-NETO, *An Introduction to thr Theory of Distributions*, 1973. rééd. Krieger Publ., Melbourne (Fla.). 198 1 / F. G FRIEDLANDER, *Introduction to the Theory of Distributions*, Cambridge Univ. Press, 1982 / 1. M. GUELFAND, G. E. SILOV, M. I. GRAEV & N. J. VILENKIN, *Les Distribution.~, 5* vol,, Dunod, Paris. 19641970 / M. HERVÉ, *Transformation de Fourier et distributions*, P.U.F., 1986 / L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators : distributions theory and Fourier analysis*, 2' éd. Springer Verlag, New York, 1990 / P. KRÉE, *Introduction aux mathématiques et à leurs applicarions fondamentales*, Dunod, Paris, 1969 / J. LÜTZEN, *The Prehistory of the Theory of Distributions*, Springer Verlag, New York-Berlin, 1982 / R. PETIT. *L 'Outil mathématique : distributions, convolutions, transformations de Fourier* et *de Laplace...*, Masson, 3' éd. 1990 / E. ROUBINE, *Distributions-signal* Eyrolles, 2' éd. 1990 / L. SCHWARTZ, *Méthodes mathématiques pour les sciences physiques, 2'* éd. Hermann, Paris, 1983 ; *Théorie des distributions*, nouv. éd. *ibid., 1984* ; *Application des distribution.~ à l'étude des particules élémentaires en mécanique quantique relativiste*, Gordon and Breach, Paris-Londres-New York. 1969.

DIVISIBILITÉ

L'étude élémentaire de la divisibilité dans l'anneau *Z* des entiers relatifs résulte de l'existence de la division euclidienne qui entraîne que cet anneau est *principal.* Les propriétés générales des anneaux principaux sont exposées dans l'article ANNEAUX COMMUTATIFS, et nous nous contenterons ici d'énumérer les principaux résultats relatifs au cas particulier qui nous occupe ici.

L'étude plus fine et plus spécifique de l'anneau *Z* (nombre de diviseurs d'un nombre donné, somme de ces diviseurs, etc.) introduit des fonctions arithmétiques multiplicatives. Les indications qui suivent sont très élémentaires, mais il est important de noter qu'un grand nombre des résultats obtenus ont été généralisés aux corps de nombres algébriques ; le dernier chapitre donne un aperçu de ces propriétés dans le cas des corps quadratiques, en renvoyant à l'article théorie des NOMBRES - Nombres algébriques pour l'exposé de la théorie sous sa forme contemporaine.



1. Propriétés élémentaires

L'anneau *Z* des entiers relatifs possède la propriété suivante de *division euclidienne :* si a et b sont deux entiers relatifs, $b \neq 0$, il existe des entiers q et r déterminés de manière unique par les conditions :

$$a = bq + r, \quad 0 \leqslant r \leqslant b - 1\,;$$

q s'appelle le *quotient* de la division de a par b et b est le *reste* de cette division. Si le reste est nul, cela signifie qu'il existe un entier q tel que $a = bq$; on dit alors que b divise a, ou que a est un multiple de h.

Dans ce qui suit, nous nous limiterons, sauf mention explicite du contraire, **aux** entiers *positif.* On écrit b a si b divise a. Cette relation de divisibilité est une relation d'ordre dans les entiers naturels ; en effet, elle est réflexive car a a, transitive car c b et b a entraînent c a, antisymétrique car a b et b a entraînent $a = b$. Cet ordre n'est pas total car deux entiers a et b ne vérifient pas obligatoirement l'une des relations $a | b$ ou $b |$ a. Un nombre $p \neq 1$ est dit *premier* s'il n'est divisible que par 1 et par lui-même.

Soient a et b deux entiers positifs ; on montre (cf. ANNEAUX COMMUTATIFS) qu'il existe un diviseur commun d de a et de b tel que les diviseurs communs de a et de b

soient exactement les diviseurs de d ; ce diviseur commun privilégié est appelé le plus grand commun diviseur (en abrégé P.G.C.D.) de a et de b et se note $d = (a,$ h). Si (a, b) = 1, c'est-à-dire si le seul diviseur commun de a et de b est 1, on dit que a et b sont *premiers entre eux.* Une condition nécessaire et suffisante pour que a et b soient premiers entre eux est qu'il existe des entiers *relatifs* x et y tels que :

$$ax + by = 1 \quad \text{(identité de Bezout).}$$

Ce résultat entraîne facilement le lemme de Gauss : si un entier c divise un produit ah et est premier avec a, alors il divise b. On en déduit le théorème fondamental de la *décomposition en facteurs premiers* : tout entier naturel a > 1 est décomposable, d'une manière unique, en un produit :

$$a = p_1^{k_1} p_2^{k_2} \dots p_n^{k_n}$$

de nombres premiers $p_1, p_2, \dots, p_n$ distincts ; les exposants k, sont des entiers $\geqslant$ 1. On peut, à partir de là, donner la règle d'obtention du P.G.C.D. de nombres ainsi décomposés : prendre les facteurs premiers communs avec les exposants les plus petits. On peut aussi bâtir sur ces décompositions la théorie du plus petit commun multiple (en abrégé P.P.C.M.) de deux nombres : prendre les facteurs premiers, communs ou non, avec leurs exposants les plus grands. Le P.P.C.M. ainsi introduit de a et b est un multiple commun m tel que les multiples communs de a et de b soient exactement les multiples de m ; il est relié au P.G.C.D. d de a et de b par la relation $md = ub$.

Indiquons enfin que les notions de P.G.C.D. et de P.P.C.M. s'étendent sans difficulté au cas de n entiers a_1, a,, ... a,.

2. Fonctions arithmétiques

Congruences

On se placera ici dans l'anneau Z des entiers relatifs. On dit que a *est congru à b* (modulo m), ce qui s'écrit a $\equiv b$ (mod m), lorsque m (u-b). Cette congruence modulo m, pour m fixé, est une relation d'équivalence (réflexive, transitive, symétrique) et permet donc de faire une partition de Z en classes (ensemble quotient par cette équivalence). Chacune de ces classes est appelée *classe résiduelle* modulo m, **et** comprend un élément et un seul a compris entre 0 et $(m - 1)$, soit $0 \leqslant a \leqslant m$ 1, tel que tout autre élément de la classe est égal à a + km. Si l'on désigne par [b],, la classe d'un entier b, il y a ainsi m classes, à savoir : [OI,,,. $[1]_m, [2]_m, \dots . [m-1]_m$. On dit que m nombres b,, $b_2, \dots .$ b,,, forment un *système complet de résidus,* modulo m, si ces nombres sont, deux à deux, non congrus modulo m ; ils correspondent donc aux m classes. La congruence modulo m étant stable dans l'addition et dans la multiplication, on peut munir l'ensemble des classes résiduelles des opérations de somme et de produit (avec [a],,, + [b], = [a $+$ b], et $[a]_m$ X [b],, = [ab],). On obtient ainsi un anneau Commutatif (Cf. ANNEAUX ET ALGÈBRES), dans lequel $[ax]_m = [ay]_m$ entraîne $[x]_m = [y]_m$ si a est premier à m ; dans le cas général, on n'aurait que $[x]_{m/d} = [y]_{m/d}$ avec d = (u, m). Un cas particulièrement intéressant est celui où $m = p$ est premier ; dans ce cas, en effet, l'anneau des classes résidueiies devient un corps, qu'on écrit en général $\mathbf{Z}/p\mathbf{Z}$ ou $\mathbf{Z}/p$. En effet, si [a], $\neq$ 0, donc a non multiple de p, c'est-à-dire a premier avec p, on peut écrire $ab + pc = 1$, d'où $[a]_p[b]_p = [1]_p$, ce qui montre l'existence de 1 / [a], = $[b]_p$. Cela rejoint d'ailleurs une propriété classique : tout anneau fini sans diviseur de 0 est un

corps. D'autre part, il est facile de voir que $[a]_m = [b]_m$ équivaut à $(a, m) = (b, m)$; on peut donc envisager les classes résiduelles premières à m qui forment, pour m donné, ce qu'on appelle un *système réduit* (par exemple, pour $m = 18$, les classes [1], [5], [7], [11], [13] et [17]). On dit alors qu'un ensemble d'entiers est un *système réduit de résidus* modulo m, si un et un seul de ces entiers appartient à chacune des classes d'un système réduit. Pour $m = p$ premier, toutes les classes sauf [0] forment un système réduit, comprenant $(p - 1)$ éléments. Dans le cas général de m quelconque, le nombre d'éléments d'un système réduit est égal à celui des nombres compris entre 1 et m et premiers à m. Ce nombre s'appelle *l'indicateur d'Euler de m*, désigné par $\varphi(m)$, et l'on peut établir les conditions nécessaires et suffisantes suivantes pour qu'un ensemble de nombres soit un système réduit de résidus modulo m : ce système comprend $\varphi(m)$ éléments ; ces éléments sont deux à deux non congrus modulo m ; chaque élément est premier à m. Il est alors évident que, si l'on multiplie chacun de ses éléments par un même facteur premier à m, on transforme un système réduit de résidus en un système réduit.

Fonctions arithmétiques classiques

La fonction φ d'Euler est une *fonction arithmétique multiplicative* ; on appelle ainsi toute fonction f définie sur les entiers naturels, et telle que $f(ab) = f(a)f(b)$ lorsque $(a, b) = 1$. On établit sur les fonctions arithmétiques multiplicatives l'important théorème suivant : si f est arithmétique multiplicative et si l'on pose :

$$F(n) = \sum_{d|n} f(d),$$

alors F est aussi multiplicative, et réciproquement.

De plus, on a :

$$f(n) = \sum_{d|n} F\left(\frac{n}{d}\right)\mu(d),$$

où :

$$\mu(d) = \begin{cases} 1, \text{ si } d = 1; \\ 0, \text{ si } d \text{ contient des facteurs carrés ;} \\ (-1)^{\nu}, \text{ si } d = p_1 p_2 \dots p_{\nu}, \text{ produit de nombres premiers distincts.} \end{cases}$$

Cette fonction μ s'appelle *fonction de Möbius* ; elle est aussi multiplicative. La relation de réciprocité liant f à F, grâce à cette fonction de Mobius, reste d'ailleurs valable dans le cas de fonctions non multiplicatives ; elle s'établit par simple calcul, et fournit la démonstration la plus simple du fait que f est multiplicative si F l'est. L'implication en sens contraire est beaucoup plus simple à établir ; elle résulte de ce que tout diviseur d'un produit de deux nombres premiers entre eux est le produit d'un diviseur de l'un par un diviseur de l'autre.

On établit donc que la fonction φ d'Euler est arithmétiquement multiplicative, soit en la mettant sous la forme :

$$\varphi(m) = m \prod_{p|m}\left(1 - \frac{1}{p}\right),$$

car c'est :

$$m - \sum \frac{m}{p_i} + \sum \frac{m}{p_i p_j} - \sum \frac{m}{p_i p_j p_k} + \dots,$$

où les p_i sont les facteurs premiers de m ; soit aussi en considérant les nombres de la forme $ax + by$ qui sont premiers à ab (système réduit de résidus, modulo ab).

Si l'on envisage alors :

$$\Phi(m) = \sum_{d|m} \varphi(d),$$

on a une fonction multiplicative qui, par simple calcul, donne $\Phi(p^\alpha) = p^\alpha$, donc $\Phi(m) = m$ pour tout m. Cela établit la formule classique :

$$\sum_{d|m} \varphi(d) = m,$$

qui peut aussi se démontrer en rangeant 1, 2, 3, ... m en classes C_d, comprenant les nombres dont le P.G.C.D. avec m est d; il y a $\varphi(m/d)$ nombres dans C_d, d'où la formule précédente, par simple décompte.

La formule de réciprocité de Mobius, appliquée à :

$$\sum_{d|m} \varphi(d) = m,$$

donne alors : $\varphi(m) = m \sum_{d|m} \frac{\mu(d)}{d}$.

Deux autres fonctions multiplicatives sont classiquement attachées à la fonction φ d'Euler ; il s'agit de $\tau(n)$ =nombre des diviseurs positifs de n (1 et n compris) et o(n) = somme de ces diviseurs. La multi-phcativité de ces deux fonctions découle encore de ce que, si a et b sont premiers entre eux, tout diviseur de ab est le produit d'un diviseur de a par un diviseur de b. Comme :

$$\tau(p^\alpha) = \alpha + 1 \quad \text{et} \quad \sigma(p^\alpha) = \frac{p^{\alpha+1} - 1}{p - 1},$$

il vient : $\tau(n) = \Pi(\alpha_i + 1)$

et
$$\sigma(n) = \prod \frac{p_i^{\alpha_i+1} - 1}{p_i - 1},$$
$$\text{où} : n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_r^{\alpha_r}.$$

À noter la relation :

$$\sum_1^n \tau(k) = \sum_1^n \left[\frac{k}{n}\right],$$

où la quantité entre crochets représente ici la partie entière de k/n.

Nombres parfaits

On appelle *nombre* parfait un nombre tel que o(n) = $2n$, et on a établi que tout nombre parfait pair s'écrit sous la forme :

$$n = 2^{p-1}(2^p - 1),$$

avec p et $(2^p - 1)$ premiers (Euclide avait déjà étudié sous cette forme les nombres parfaits). On ne sait pas, actuellement, s'il y a ou non des nombres parfaits impairs. Les nombres parfaits pairs sont donc liés aux nombres premiers de la forme $2^p - 1$. Ces nombres sont appelés *nombres de Mersenne* (Mersenne affirma en 1644 que, jusqu'à p = 257, $2^p - 1$ était premier seulement sip = 2, 3, 5, 7, 13, 17, 19, 31, 67, 127 et 257). On a établi, depuis, que cette liste contenait des erreurs et des omissions : jusqu'à 5 000, $(2^p \quad 1)$ est premier pourp = 2, 3, 5, 7, 13, 17, 19, 31, 61, 89, 107, 127, 521, 607, 1279, 2 203, 2281, 3 217, 4253 et 4423. Fermat, de son côté, avait conjecturé que les *nombres de Fermat* :

$$F_n = 2^{2^n} + 1$$

étaient tous premiers. Euler en 1732 établit que F_5 était composé (divisible par 641), Legendre en 1780 établit que F_6 est divisible par 274 177. On a établi, depuis, que F_n est non premier pour $7 \leqslant n \leqslant 16$ et n = 18, 23, 36, 38, 39, 55, 63, 73, ; et on ne connaît explicitement aucun F_n premier pour $n > 4$. Les nombres de Fermat premiers jouent un rôle essentiel dans !a recherche des polygones réguliers de m côtés que l'on peut construire avec la règle et le compas (Gauss, âgé de dix-sept ans, établit qu'une condition nécessaire et suffisante pour que cette construction soit possible est que m soit un produit de nombres de Fermat premiers distincts).

3. Théorèmes classiques

Théorème d'Euler-Fermat

Euler établit, en 1760, le théorème d'Euler-Fermat, suivant lequel, si m est entier naturel, et (a, m) = 1, on a :

$$a^{\varphi(m)} \equiv 1 \pmod m.$$

En effet, un système réduit de résidus modulo m, soit $r_1, r_2, \ldots r_{\varphi(m)}$ est transformé, on l'a vu, en un système réduit par multiplication par a ; donc $ar_1, ar_2, \ldots, ar_{\varphi(m)}$ sont, modulo m, égaux à une permutation de $r_1, r_2, \ldots r_{\varphi(m)}$ d'où le produit :

$$a^{\varphi(m)} r_1 r_2 \, r_{\varphi(m)} \equiv r_1 r_2 \ldots r_{\varphi(m)} \pmod m,$$

que l'on peut simplifier par $r_1 \, r_2 \, r\varphi_{(m)}$ premiers à **m.** D'où la formule d'Euler-Fermat. Fermat avait établi en 1736 ce théorème dans le cas particulier de $m = p$ premier. Il s'agit du « petit théorème de Fermat », suivant lequel :

$$a^{p-1} \equiv 1 \pmod p,$$

si a n'est pas multiple dep. On l'écrit, sans condition sur a, sous la forme $a^p \equiv a (\mathrm{mod}\ p)$. On remarquera que la réciproque de ce théorème n'a pas lieu (par exemple m = 561 = 3 X 11 X 17 vérifie, pour tout a premier à m, $a^{560} \equiv 1$). On a d'autre part établi (Beeger en 195 1) qu'il existe une infinité d'entiers n pairs tels que 2^n 2 est divisible par n (le plus petit de ces nombres étant 161 038).

Théorème de Wilson

Le théorème de Wilson énonce que :

$$(p-1)! \equiv -1 \pmod p,$$

pour tout p premier (théorème publié en 1770 par Waring et démontré par Lagrange en 1771). Supposant p impair et développant :

$$(x-1)(x-2)\ldots(x-p+1) = x^{p-1} - A_1 x^{p-2} + \quad + \quad A_{p-1},$$

Lagrange obtient, après avoir multiplié par x puis changé x en (x- 1), une identité permettant d'établir que p A,, p A,, ..., $p \mid A_{p\ 2}$ et $(p-1)\, A_{p-1} \equiv 1$ (modp).Ona donc $A_{p-1} \equiv -1$ (modp), ce qui établit le théorème, car $A_{p-1} = (p-1)\,!$. On a de plus :

$$x^{p-1} - 1 \equiv (x-1)(x-2) \ldots (x-p+1) \pmod p,$$

ce qui, pour x premier à p, donne le petit théorème de Fermat. On verra une autre démonstration du théorème de Wilson, dans le paragraphe des résidus quadratiques.

Racines primitives

La notion de racine primitive modulo m, est liée à la formule d'Euler :

$$a^{\varphi(m)} \equiv 1 \pmod m.$$

Soit en effet k le plus petit exposant pour lequel $a^p \equiv 1$ (mod **m).** On dit que a appartient à l'exposant k (mod m), ***ou*** que k est l'ordre de ***a,*** modulo **m.** Il s'ensuit que k divise tout x tel que $a^x \equiv 1$ (mod ***m)*** ; en particulier, k φ ***(m),*** et on dit que a est une racine primitive modulo m si l'ordre de ***a*** est ***<p(m),*** c'est-à-dire si $\varphi(m)$ est effectivement le plus petit exposant pour lequel $a^p \equiv 1$ (mod ***m).*** Cette définition montre que tout entier congru, modulo m, ***à*** une racine primitive, en est également une. Ces racines primitives ont l'importante propriété que, pour chacune d'elle, a par exemple, les nombres a, $a^2, a^3, \ldots a^{\varphi(m)}$ forment un système réduit de résidus. On démontre que tout p premier possède des racines primitives ; il y en a exactement

$\varphi(p-1)$ distinctes entre elles, modulo p. Cela découle d'un théorème assez surprenant suivant lequel, si ***p*** est premier, et ***d*** $(p-1)$, il y a exactement $\varphi(d)$ nombres non congrus 2 à 2 et dont l'ordre est ***d*** (modulo ***p)*** ; le nombre des nombres d'ordre ***d*** (modulo ***p)*** ne dépend donc pas de p, dès que ***d*** divise $(p \quad 1)$. Pour l'existence des racines primitives modulo m, avec m non premier, l'étude, plus délicate, fut faite par Gauss, qui établit que ces racines primitives n'existent que si $m = 2$, 4 ou p^k ou $2p^k$ (p premier impair quelconque).

4. Résidus quadratiques

Résidus et non-résidus

Un nombre a premier à m est dit ***résidu quadratiqur*** de m, si $x^2 \equiv a$ (mod ***m)*** a des solutions entières en x ; sinon a est dit ***non-résidu quadratique*** (avec toujours la condition ***a*** premier à m). Dans le cas où $m = p$ premier, il est facile de voir qu'il existe, modulo p, $(p-1)/2$ résidus quadratiques et $(p-1)/2$ non-résidus ; en effet, $1^2, 2^2, \ldots, (p-1)^2$ donnent, modulo p, $(p-1)/2$ classes résiduelles différentes ; car $(p-q)^2 \equiv q^2$ et $a^2-b^2 = (a-h)(a+b) \equiv 0$ (modp) si a et b sont au plus égaux à $(p-1)/2$. Par exemple, pour ***p*** = 11, on a les résidus quadratiques 1, 3, 4, 5 et 9. On peut établir aisément, pour m quelconque, que le produit de deux résidus quadratiques de m est un résidu ; car $x^2 \equiv a$ et $y^2 \equiv b$ entraîne $(xy)^2 \equiv ab$ (mod m). Dans le cas où $m = p$ premier, le produit d'un résidu par un non-résidu est un non-résidu et le produit de deux non-résidus est un résidu ; il suffit pour cela d'envisager a, 2 ***a, 3 a, ...*** $(p \quad 1)a$, qui sont non congrus modulo p, donc forment un système complet (mod ***p).*** Il y a donc $(p-1)/2$ résidus et $(p-1)/2$ non-résidus, quel que soit ***a*** premier à p, et, si a est résidu, les $(p-1)/2$ résidus proviennent du produit par a des résidus quadratiques de p, donc les $(p \quad 1)/2$ non-résidus correspondent aux produits de ***a*** par les non-résidus de ***p.*** Si ***a*** est non-résidu, les $(p-1)/2$ non-résidus proviennent donc du produit de ***a*** par les résidus quadratiques de ***p,*** donc les $(p-1)/2$ résidus correspondent aux produits de a par les non-résidus de p. Pour m quelconque, cependant, on peut avoir le produit de deux non-résidus qui soit un non-résidu : par exemple, pour m = ***45,*** les résidus quadratiques sont 1, 4, 16, 19, 31, 34, et 2 X 7 $\equiv$ 14. Un critère d'Euler établit que, pour ***p*** premier différent de 2, a est résidu ou non-résidu quadratique de ***p*** suivant que, respectivement,

$$a^{(p-1)/2} \equiv 1 \pmod p$$

ou :
$$a^{(p-1)/2} \equiv -1 \pmod p ;$$

on ne peut avoir que l'une ou l'autre de ces congruences puisque $a^{p-1} \equiv 1$ (modp). On peut, à partir de ce critère, retrouver les théorèmes concernant les produits de résidus ou non-résidus (mod ***p).***

Loi de réciprocité

Legendre a introduit un symbole qui porte son nom : pour ***p*** premier,

$$\left(\frac{a}{p}\right) = +1,$$

si ***a*** est résidu ;

et
$$\left(\frac{a}{p}\right) = -1,$$

si ***a*** est un non-résidu. On a donc :

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right).$$

Ce symbole permet d'exprimer un important théorème connu sous le nom de ***loi de réciprocité quadratique.*** Cette loi fut prouvée par Euler en 1783, retrouvée par Legendre en 1785 et mise au point par Gauss en 1808 ; elle s'écrit, pour deux premiers impairs distincts ***p*** et q, ***sous*** la forme :

$$\left(\frac{p}{4}\right)\left(\frac{q}{P}\right)=(-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}$$

En d'autres termes les congruences $x^2 \equiv p$ (mod ***q)*** et $y^2 \equiv$ ***q*** (mod ***p)*** sont résolubles ensemble, ou non, sauf si ***p*** $\equiv$ ***q*** $\equiv 3$ (mod 4), auquel cas une et une seule de ces équations est résoluble. Le symbole de Legendre a été étendu par Jacobi, qui définit

$$\left(\frac{a}{p_1p_2\cdots p_n}\right)=\left(\frac{a}{p_1}\right)\left(\frac{a}{p_2}\right)\cdots\left(\frac{a}{p_n}\right),$$

les nombres premiers p_i étant distincts ou non. Ce symbole, toutefois, a l'inconvénient d'être égal à + 1 sans que ***a*** soit toujours résidu quadratique modulo p_1 p_2 p_n ; par exemple :

$$\left(\frac{a}{9}\right)=\left(\frac{a}{3}\right)\left(\frac{a}{3}\right)$$

vaut toujours $+$ 1. On peut cependant établir, pour ***a*** et b premiers entre eux, que

$$\left(\frac{a}{b}\right)=-1$$

entraîne que ***a*** est non-résidu quadratique modulo b. Signalons enfin que le théorème de Wilson : $(p-1)! \equiv -1$ (modp), établi antérieurement, se démontre à partir des résidus : pour ***p*** premier impair et pour ***a*** non multiple de p, on peut établir que :

$$(p-1)!\equiv-\left(\frac{a}{b}\right)a^{\frac{p-1}{2}}\pmod{p};$$

il suffit, pour cela, d'associer les couples x et $(p-x)$ et de voir que, deux à deux, on a $x'x'' \equiv$ ***a*** (mod ***p)*** ou bien, si ***a*** est résidu, $x_1^2 \equiv$ ***a*** et $(p-x_1)x_1 \equiv$ ***-a*** (modp) et $x'x'' \equiv$ ***a*** (mod ***p)*** pour les autres. Les résidus quadratiques sont utilisés en particulier dans la théorie des corps quadratiques, dans la factorisation des nombres (par exemple, si N = $x^2 + ky^2$, avec x et y premiers entre eux, -k doit être résidu quadratique de tous les facteurs premiers de N, ce qui facilite la factorisation), et aussi dans la recherche des carrés parfaits dans le corps $\mathbf{Q}_p$ des nombresp-adiques. **À** côté de la recherche directe, par développements de Hensel, l'introduction du symbole de Hilbert ***(a,*** β), égal à $+1$ ou -1 suivant que $\alpha x^2 + \beta y^2 - z^2 = 0$ est résoluble ou non dans $\mathbf{Q}_p$, conduit à la caractérisation des carrés ***a*** par le fait que ***(a,*** β) = 1 pour tout β.

5. Divisibilité dans les corps quadratiques

On ne donnera ici qu'un aperçu de la théorie de la divisibilité dans les corps quadratiques. Si l'on considère les nombres de la forme :

$$\frac{u+v\sqrt{d}}{w}$$

ou ***d*** est entier non carré parfait, et u, v, w entiers relatifs (avec $w \geqslant 1$), on définit un corps, appelé ***corps quadratique Q*** $(\sqrt{d})$. Dans ce corps, on appelle ***entiers*** les éléments qui vérifient une équation du type $\alpha^2 + a_1\alpha + a_2 = 0$, ***a,*** et a_2 étant des entiers ; et on démontre que ces entiers sont donnés par les formules :

$$a+b\sqrt{d},\ \text{si}\quad d\equiv 2\ \text{ou}\quad 3\pmod 4$$

$$\text{et}\qquad a+b\,\frac{\sqrt{d}-1}{2},\ \text{si}\quad d\equiv 1\pmod 4.$$

Ces entiers forment un sous-anneau de Q ($\sqrt{d}$), et on peut définir dans cet anneau la divisibilité, compliquée par le fait qu'il existe d'autres *unités* que + 1 ou 1. Une unité quadratique est en effet racine d'une équation :

$$\alpha^2 + a_1 \alpha \pm 1 = 0$$

Il y a une infinité d'unités dans Q ($\sqrt{d}$) pour $d \geqslant 2$ et, pour $d \leqslant -1$, il n'y en a pas d'autre possible que $1, -1$; $i, -i$ et les racines troisièmes de l'unité $j, j^2, -j$ et $-j^2$. Une unité divise tout entier; on définira donc les nombres premiers comme étant ceux qui ne sont divisibles que par eux ou par les unités du corps. De même, a et *b* seront dits premiers entre eux si leurs seuls diviseurs communs sont les unités ; on écrit encore $(a, b) = 1$ mais c'est un symbole car 1 n'est plus le P.G.C.D. au sens ordinaire. Sans entrer dans le détail, signalons qu'alors le théorème de Gauss (a hc et (a, *b*) = 1 entraînent a c) peut avoir lieu, ou ne pas avoir lieu, suivant d. Lorsque ce théorème a lieu, Q ($\sqrt{d}$) est appelé corps quadratique *simple* ; en découle une décomposition unique en facteurs premiers (à des facteurs unités près). Par exemple, il en est ainsi pour $d = -1$, $d = 2$, $d = -3$, mais pas pour $d =$ 5 ou $d = 10$ (on a par exemple :

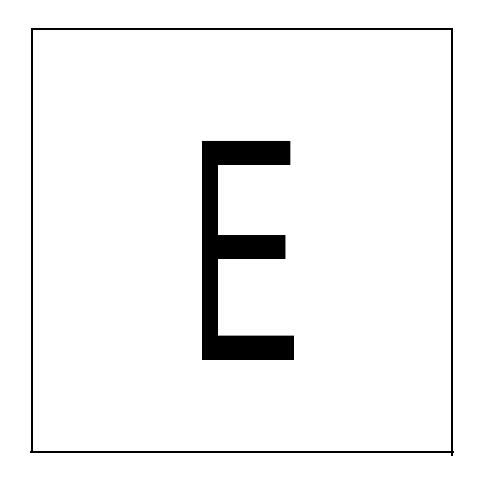
$$6 = 2 \times 3 = (4 + \sqrt{10})(4 - \sqrt{10})$$

et on vérifie qu'il n'y a pas d'unités permettant de passer d'une décomposition à l'autre). Les seuls cas quadratiques simples, pour d < 0, sont les cas où $-d$ = 1, 2, 3, 7, 11, 19, 43, 67, 163 (résultat de Stark et Baker en 1966 ; avant eux on avait établi qu'il en existait peut-être encore un, avec $d > 5 \times 10^9$). Une autre notion peut s'étendre à Q ($\sqrt{d}$) ; il s'agit de la division euclidienne, qui fait intervenir les normes des nombres quadratiques, soit N (α) = $\alpha \bar{\alpha}$ (où $\bar{\alpha}$ est le conjugué de α). Q ($\sqrt{d}$) est dit *euclidien* si, pour tout α_1 et α_2 entiers, on peut écrire $\alpha_1 = \beta\alpha_2 + \gamma$ avec $N(\gamma) < N(\alpha_2)$. La propriété Q ($\sqrt{d}$) « euclidien » entraîne évidemment Q ($\sqrt{d}$) « simple », mais pas réciproquement. On a démontré que les seuls cas euclidiens correspondent à $d = -11, -7, -3, -2, -1, 2, 3, 5, 6, 7,$ 11, 13, 17, 19, 21, 29, 33, 37, 41, 57 et 73. Le cas $d = -1$ est le cas, bien connu, des entiers de Gauss a + *bi* où *a* et *b* sont des entiers relatifs ordinaires.

MARCEL DAVID

Bibliographie

Z. J. BOREVITCH & I. R. CHAFAREVITCH, *Théorie des nombres*, trad. J. L. Verley, Gauthier-Villars, Paris, 1967 / G. H. HARDY & E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Oxford Univ., New York, 1979 / K. IRELAND & M. ROSEN, *A Classical Introduction to Modern Number Theory*, Springer-Verlag, New York, 2ᵉ éd. 1990 / E. LUCAS, *Théorie des nombres : le calcul des nombres entiers, le calcul des nombres rationnels, la divisibilité arithmétique*, 1958, repr. A. Blauchard, 1979 / T. NAGELL, *Introduction to Number Theory*, Chelsea Publ., New York, 1981 / I. NIVEN & H. S. ZUCKERMANN, *An Introduction to the Theory of Numbers*, Wiley, New York, 1980 / H. N. SHAPIRO, *Introduction to the Theory of Numbers*, Wiley, New York, 1983 / H. M. STARK, *An Introduction to Number Theory*, M.I.T. Press, Cambridge (Mass.), 1978



ENSEMBLES THÉORIE ÉLÉMENTAIRE DES

Toute pensée formalisée s'exprime de nos jours dans le langage de la théorie des ensembles, qui a ainsi envahi toutes les disciplines, sciences humaines comprises. Dès l'école primaire, l'enfant apprend à classer des objets suivant leur forme, leur couleur, leur taille, à établir entre eux des correspondances, préambules à des manipulations plus abstraites.

La théorie élémentaire des ensembles fait partie du bagage culturel minimal de l'homme contemporain.

L'algèbre des ensembles n'est pas non plus étrangère aux progrès de la technique, ne serait-ce que parce qu'elle joue un grand rôle en informatique dans la conception et la construction des ordinateurs et des logiciels ; elle intervient aussi pour une large part dans l'organisation de l'information, les techniques de gestion, les études de marché.

Les règles générales de la logique d'Aristote ont été, depuis le Moyen Âge, l'objet de recherches visant à dégager un « langage universel » commun à la logique formelle et à l'algèbre. C'est Leibniz qui formule le premier la demande, sous-jacente chez Raymond Lulle, d'un « alphabet des pensées humaines » permettant de réduire à un algorithme symbolique le raisonnement déductif.

Pour Leibniz, sa « caractéristique universelle » est un langage formalisé, pure combinaison de signes, dont les théorèmes ou les propositions se déduiraient de manière purement mécanique (*calculus ratiocinator*). En ce sens, il est le précurseur de la théorie élémentaire des ensembles.

Inédits et inconnus jusqu'au début du XIXe siècle, les travaux de Leibniz sont repris par les logiciens de l'école anglaise. C'est G. Boole qui est le véritable créateur du calcul direct sur les parties d'un ensemble. A. de Morgan (1858), W. S. Jevons (1864) et C. S. Pierce (1867) abordent l'étude des *relations* sur un ensemble.

En 1890, E. Schröder, dans un ouvrage monumental, fait la synthèse des travaux de ses devanciers. À la fin du XIXe siècle, les recherches de Frege et de Peano s'articulent avec les problèmes des fondements des mathématiques.

On ne donnera pas dans cet article une construction formelle et rigoureuse de la théorie des ensembles, mais on essayera, à partir de quelques notions premières considérées comme intuitives, d'indiquer les résultats les plus élémentaires.

E. U.



1. Calcul booléen

Les ensembles

Leibniz, philosophe et mathématicien (1646-1716), recherche un système qui lui

permette de formaliser le langage et la pensée. Pour lui, un langage formalisé doit être une pure combinaison de signes, dont seul importe l'enchaînement, de sorte qu'une machine serait capable de fournir tous les théorèmes et que toutes les controverses se résoudraient par un simple calcul.

Leibniz, n'arrivant pas à exploiter toutes ses idées, dont certaines auraient pu le conduire à de meilleures conclusions, échoue dans sa tentative. Après lui, pendant tout le XVIIIe siècle et au début du XIXe, d'autres auteurs ébauchent des tentatives semblables sans arriver à avancer plus que Leibniz. À cette époque, leurs travaux, de même que ceux de Leibniz, ne sont pas connus et n'ont qu'un très faible retentissement ; chacun ignore les travaux de ses prédécesseurs.

C'est dans les mêmes conditions que le mathématicien anglais George Boole (1815-1864) va travailler. Boole peut être considéré comme le véritable créateur de la logique contemporaine. Son ambition est de formaliser la logique en s'inspirant des méthodes de l'analyse et de l'algèbre : « Que l'on donne des formes existantes de l'analyse une interprétation quantitative n'est que le résultat des circonstances dans lesquelles elles furent établies et ne doit pas être érigé en condition universelle de l'analyse. C'est sur le fondement de ce principe général que je me propose d'établir le calcul logique et que je lui réclame une place parmi les formes reconnues de l'analyse mathématique, sans égard au fait qu'en son objet comme en ses instruments il doive actuellement demeurer en dehors d'elle » (*Thr Muthematical Analysis oj Logic*, 1847).

Ce sont les travaux de Boole qui ont donné naissance à ce qu'on appelle aujourd'hui l'algèbre de Boole (ou calcul booléen), en donnant un point de départ au calcul des propositions et à l'algèbre des ensembles.

Notion d'ensemble

Trois mots ou symboles seront constamment utilisés dans ce qui suit : « ensemble », « élément », « $\in$ » (qui se lit « est élément de » ou « appartient à »). Il est impossible de définir ces mots. En effet on pourrait dire : « Un ensemble est une collection d'objets », ou encore comme Cantor (dans *Gesammelte Abhanlungen*) : « Par ensemble, on entend un groupement en un tout d'objets bien distincts de notre intuition ou de notre pensée. » Dans ce cas, on ne fait que déplacer le problème et il reste à définir les mots « collection », « groupement », « objets ». La situation est tout à fait comparable à celle qu'on rencontre lorsqu'on veut reconstruire la géométrie ; il n'est pas possible de définir les mots « point », « droite », « plan ». Pour les objets correspondants, on indique leurs propriétés et les règles d'utilisation : ce sont les axiomes.

Un ensemble est constitué d'éléments. Une image intuitive d'un ensemble est donnée par une collection d'objets, un groupement d'objets. Un élément d'un ensemble peut être soit un animal, soit un objet, soit un être mathématique, soit lui-même un ensemble. 11 existe en mathématique de nombreux exemples d'ensembles d'ensembles.

Dans la suite, on représentera les éléments et les ensembles par des lettres de l'alphabet. Souvent les éléments sont désignés par des lettres minuscules et les ensembles par des lettres majuscules. Rien n'oblige à respecter cette convention et cela devient impossible dans le cas des ensembles d'ensembles.

Pour exprimer que a est élément de l'ensemble E, que a appartient à l'ensemble E, on écrit :

$a \in E.$

Pour exprimer que a n'est pas élément de l'ensemble E, que a n'appartient pas à l'ensemble E, on écrit :

$a \notin E.$

L'expression $a \notin E$ est la négation de l'expression $a \in E$. Par suite, en vertu du principe du tiers exclu en logique et des relations qui existent entre une proposition et sa négation, il est possible de dire que, pour *a* donné et E donné, seule l'une de ces deux expressions est vraie et alors sa négation est fausse.

Il faut ajouter à cela quelques règles propres à la théorie des ensembles :

Règle 1. Un ensemble est parfaitement défini par la connaissance des éléments qui le constituent, ou encore : Deux ensembles sont égaux si, et seulement si, ils sont constitués par les mêmes éléments.

Pour se donner un ensemble il suffit donc de se donner la liste de ses éléments. Exemples : l'ensemble constitué par les lettres *a*, e, i, o, u ; on le note habituellement : { *a*, e, i, o, u }. L'ensemble constitué par les nombres 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 sera noté de la même manière : { 0, 1, 2, 3, 4, 5, 6, 7, 8. 9, 10, 11, 12 }.

Mais il y a une deuxième manière de préciser un ensemble, en les construisant à partir d'ensembles déjà connus. Ainsi, si E désigne l'ensemble des êtres humains de nationalité française et si on considère l'expression « x est inscrit sur les listes électorales », cette expression est vraie pour certains Français, fausse pour d'autres. Elle va permettre de distinguer, dans l'ensemble des Français, ceux pour lesquels cette expression est vraie, et de définir ainsi un nouvel ensemble. Cet ensemble sera noté :

{ $x \in E$; x est inscrit sur les listes électorales }.

De même, dans l'ensemble A des lettres de l'alphabet français, on peut considérer celles pour qui l'expression : « □ est une voyelle » est vraie. On définit un ensemble noté :

{□ ∈ A; □ est une voyelle}.

On aurait aussi un autre ensemble :

{ □ ∈ A ; □ n'est pas une voyelle }.

De même :

{ $x \in E$; x n'est pas inscrit sur les listes électorales }.

Admettre qu'on peut ainsi définir des ensembles revient à admettre le principe (ou règle) suivant :

Règle 2. Quand on a un ensemble A et une propriété P, on peut définir un ensemble B dont les éléments sont les éléments de A ayant la propriété P. Si on note $P(x)$ l'expression « x a la propriété P », l'ensemble B peut s'écrire :

$\{ x \in A ; \; \mathbf{P}(x) \},$

ou, en vertu de l'égalité des deux ensembles, définie par la règle 1 et qui se traduit par le signe « = » :

$B = \{ x \in A ; \; \mathbf{P}(x) \}.$

Cette règle impose une définition. En effet l'ensemble B, défini ci-dessus, est tel que tous ses éléments sont des éléments de l'ensemble A. On dit que B est un sous-***ensemble*** de A, une ***partie*** de A, ou encore que B est inclus dans A ou contenu dans A. On traduit cela par le symbole : « ⊂ » ; ainsi « B est un sous-ensemble de A » s'écrira :

$B \subset A.$

Une conséquence de la règle 2 est qu'à partir d'un ensemble A quelconque on peut définir un ensemble particulier, qu'on appellera l'*ensemble vide,* de la manière suivante : Il suffit de considérer parmi les éléments de A ceux pour lesquels la propriété « être différent d'eux-mêmes » est vraie :

$$\{x \in A ; \quad x \neq x\}.$$

On obtient ainsi un ensemble qui n'a aucun élément, appelé l'ensemble vide et noté : « $\varnothing$ ».

Paradoxe de Russell

En 1905, Bertrand Russell montre que la notion d'« ensemble des ensembles qui ne sont pas éléments d'eux-mêmes » est contradictoire. La mise en évidence de ce résultat peut se faire de la manière suivante : à première vue les ensembles peuvent se partager en deux classes, la classe de ceux qui sont éléments d'eux-memes, classe de ceux pour lesquels l'expression X $\in$ X est vraie ; la classe de ceux qui ne sont pas éléments d'eux-mêmes, ceux pour lesquels l'expression X $\in$ X est fausse ou X $\notin$ X est vraie. Désignons par A l'ensemble de tous les ensembles qui ne sont pas éléments d'eux-mêmes. Laquelle des deux expressions A $\in$ A ou A $\notin$ A est-elle vraie ?

Supposons que A $\in$ A soit vraie. Mais alors A ne peut pas être un élément de A, puisque par définition les éléments A sont des ensembles qui ne sont pas éléments d'eux-mêmes. L'expression A $\notin$ A doit être vraie aussi.

Supposons que A $\notin$ A soit vraie. Mais alors, d'après la définition de A qui est constitué par les ensembles qui ne sont pas éléments d'eux-mêmes, on doit avoir A $\in$ A vraie. Dans un cas comme dans l'autre, on aboutit à une contradiction. Il sera donc interdit de parler de « l'ensemble des ensembles qui ne sont pas éléments d'eux-mêmes », et, pour les mêmes raisons, de l'ensemble de tous les ensembles.

Ce paradoxe est à rapprocher du paradoxe célèbre du Menteur, où le problème est de savoir si l'homme qui dit : « Je mens » dit ou non la vérité en prononçant ces paroles.

Représentation graphique

Il est **souvent** commode de représenter un élément par un point du plan, et un ensemble par l'intérieur d'une courbe fermée. Ainsi la figure 1 représente un ensemble A, *a* n'est pas élément de A, *b* et *c* sont éléments de A.

Les sous-ensembles sont alors représentés comme des portions de l'ensem-

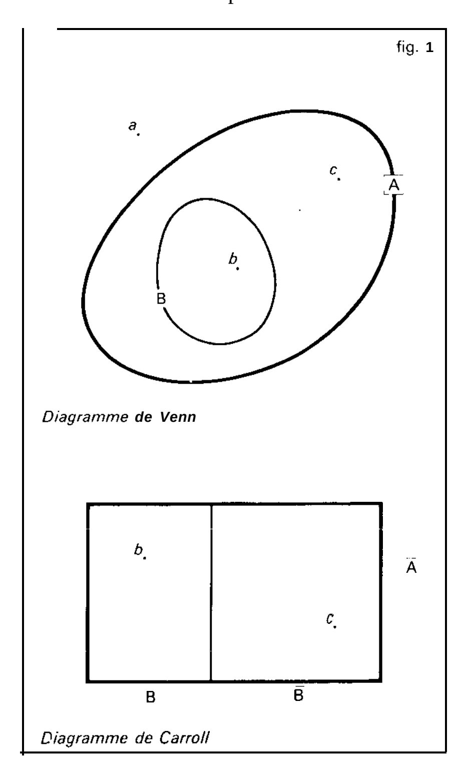


Diagramme de Venn

Diagramme de Carroll

ble. Ainsi B est une partie ou sous-ensemble de A. Ce sont les diagrammes de Venn.

Lewis Carroll propose une présentation analogue, mais l'ensemble A est représenté par un rectangle, un sous-ensemble B étant obtenu par partage du rectangle en deux par un segment de droite. Cette présentation a l'avantage de conserver une symétrie entre le sous-ensemble B et le sous-ensemble complémentaire $\bar{B}$ constitué par les éléments de A qui ne sont pas dans B (fig. 1).

De toute manière, ces représentations ne sont que des images, et il y a au moins autant de différence entre ces images et les êtres mathématiques qu'elles représentent qu'entre un schéma et l'objet représenté par ce schéma vu qu'entre l'écriture d'un mot en français et la signification de ce mot. Mais de même que les schémas permettent de se représenter les objets et servent à penser et à raisonner, de même les figures proposées ci-dessus peuvent apporter une aide importante aux raisonnements.

Ensemble des parties d'un ensemble

Un ensemble A est *partie ou sous-ensemble* de E si tous les éléments de A sont des éléments de E.

Pour un ensemble E ayant trois éléments désignés par a, b, c, il est facile d'énumérer toutes ses parties. Il y a d'abord E lui-même, qui répond à la condition ci-dessus : E C E. Ensuite les parties ayant deux éléments, ce sont : $\{a, b\}$, $\{a, c\}$, $\{b, c\}$. Ensuite les parties ayant un élément : $\{a\}$, $\{b\}$, $\{c\}$. Enfin la partie vide : $\emptyset$. Il est facile de voir que l'ensemble vide est une partie de E soit en remarquant qu'il vérifie la définition ci-dessus, soit en remarquant qu'il peut être défini de la manière suivante :

$$\emptyset = \{x \in E, x \notin E\}$$

L'ensemble des parties de E sera donc un ensemble de 8 éléments :

$$\{\{a, b, c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a\}, \{b\}, \{c\}, \emptyset\}.$$

Pour un ensemble quelconque E, on peut être amené à considérer des parties de E et l'ensemble des parties de E. Nous admettrons l'existence de ce nouvel ensemble, dont la définition pourrait être : L'ensemble des parties d'un ensemble E est un ensemble dont les éléments sont les sous-ensembles ou parties de E. Cet ensemble est noté : $\mathcal{P}$ (E). Ainsi :

$$X \subset E \Leftrightarrow X \in \mathcal{P}(E).$$

Pour l'ensemble $\{a, b, c\}$, on aura :

$$\mathcal{P}(\{a, b, c\}) = \{\{a, b, c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a\}, \{b\}, \{c\}, \emptyset\}.$$

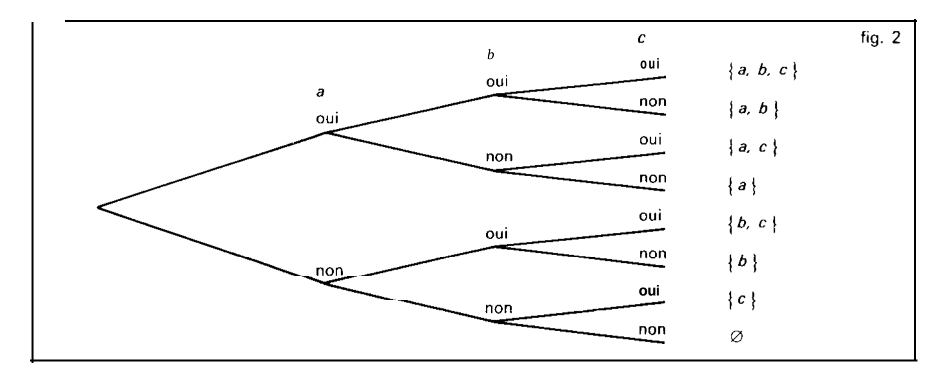
De même :

$$\mathcal{P}(\{a, b\}) = \{\{a, b\}, \{a\}, \{b\}, \emptyset\};$$
$$\mathcal{P}(\{a\}) = \{\{a\}, \emptyset\}.$$

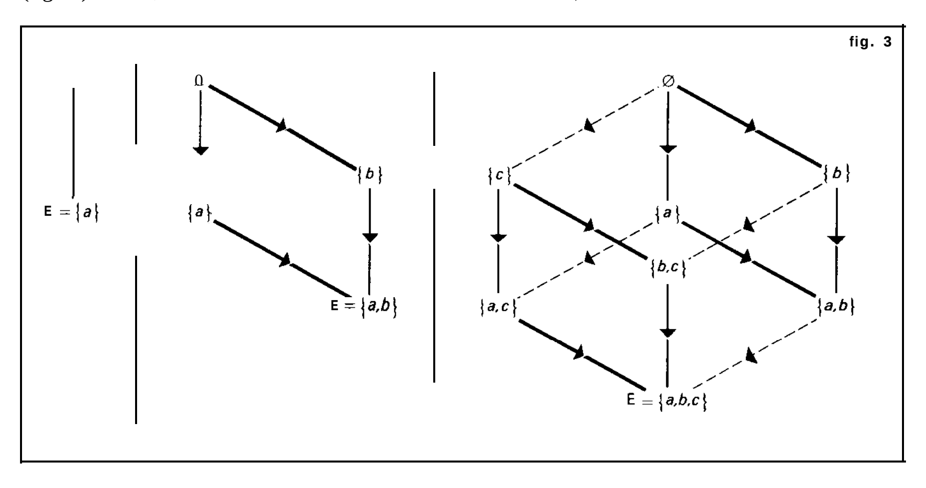
Si l'ensemble E a un élément, $\mathcal{P}$ (E) a deux éléments ; si E a deux éléments, $\mathcal{P}$ (E) en a quatre ; si E a trois éléments, $\mathcal{P}$ (E) en a huit ; on peut généraliser et voir que si l'ensemble E a n éléments, l'ensemble $\mathcal{P}$ (E) en a 2^n. Cette règle peut apparaître à l'aide d'un arbre de choix tel que celui qui est représenté sur la figure 2 dans le cas d'un ensemble à trois éléments : a, b, c.

On voit ici que, chaque fois qu'on ajoute un élément, il y aura dédoublement, donc multiplication par 2 du nombre d'éléments. On démontre alors le résultat par récurrence.

On peut aussi organiser l'ensemble des parties d'un ensemble sur un schéma en



reliant deux parties par une flèche lorsqu'on passe de l'une à l'autre par adjonction d'un élément à la première (fig. 3). Ainsi, si l'ensemble a un seul élément a, il a deux parties : ∅ et { a }. Pour E = { a, b }, on trouve quatre parties : ∅, { a }, { b }, { a, b } ; remarquons qu'on passe des deux premières aux deux dernières par adjonction de l'élément b. Pour E = { a, b, c }, on aurait un schéma obtenu par dédoublement du précédent.



On pourrait continuer ainsi pour des ensembles à 4, 5, 6, éléments. On a ici un schéma de l'organisation par la relation « est sous-ensemble de » de l'ensemble des parties d'un ensemble, une partie étant sous-ensemble d'une autre s'il existe un chemin dans le sens des flèches allant de la première à la deuxième (cf. ensembles ORDONNÉS)

Algèbre des ensembles

Pour ce qui suit, il est commode de supposer donné un ensemble E et de ne considérer que des sous-ensembles de E, donc des éléments de 𝒫 (E). On appelle quelquefois *référentiel* cet ensemble *fixe* auquel on se réfère. Cela permet de définir des opérations dans 𝒫 (E) : intersection, union, etc.

Intersection de deux ensembles

Si A et B sont deux ensembles, il peut être commode dans certains cas de considérer

l'ensemble des éléments communs. Ainsi, dans l'ensemble des Français, si A désigne l'ensemble des Français inscrits sur les listes électorales en 1970 et si B désigne l'ensemble des Français habitant la ville de Paris pendant l'année 1970, il peut être intéressant, dans certains cas, de pouvoir considérer l'ensemble des Français inscrits sur les listes électorales et habitant la ville de Paris en 1970. Un tel ensemble, constitué par les éléments communs à l'ensemble A et à l'ensemble B, sera appelé *l'intersection de* A *et de* B. On le note $A \cap B$, ce qui se lit « intersection de A et de B ».

Ainsi, si, dans l'ensemble des mots français, A désigne l'ensemble des mots dont la première lettre est un « a », B désigne l'ensemble des mots dont la dernière lettre est un « p », $A \cap B$ désigne alors l'ensemble des mots dont la première lettre est un « a » *et* la dernière est un « p ».

On peut donner une définition formelle de l'intersection :

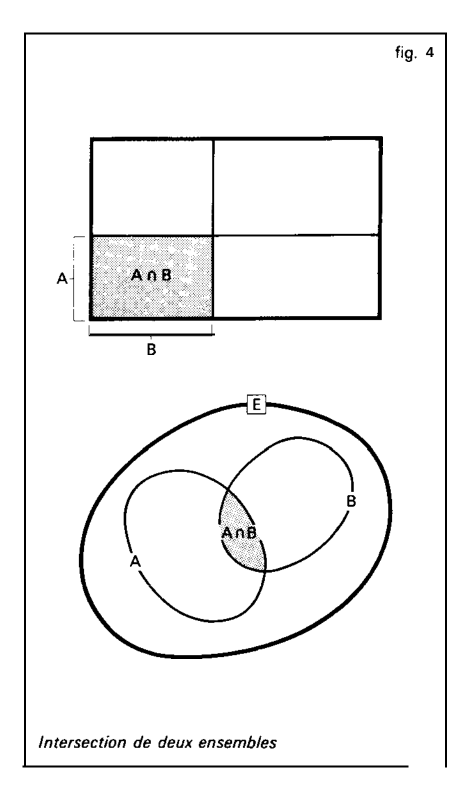
$$A \cap B = \{ x \in E ; x \in A \text{ et } x \in B \}.$$

L'intersection de A et de B est l'ensemble des éléments de E qui sont éléments de A *et* éléments de B (fig. 4). Dans cette définition, un mot est important, c'est le mot « et ». La définition est fondée sur la conjonction de deux propriétés : l'appartenance à A, l'appartenance à B. Toutes les propriétés de l'intersection sont des traductions, en langage des ensembles, des propriétés du mot « et » dans le langage courant et en logique des propositions.

Voici un autre exemple d'intersection :

$$A = \{ 1, 2, 3, 6, 9, 18 \},$$
$$B = \{ 1, 3, 5, 9, 15, 45 \},$$
$$A \cap B = \{ 1, 3, 9 \},$$

où A est l'ensemble des diviseurs de 18, B est l'ensemble des diviseurs de 45, et $A \cap B$



Intersection de deux ensembles

est l'ensemble des diviseurs communs à 18 et à 45.

Dans le cas où l'intersection de A et de B est l'ensemble vide, on dit que A et B sont *disjoints.*

Parmi les propriétés de l'intersection, on peut signaler :

(a) $A \cap B \subset A$ et $A \cap B \subset B$,

i.e. l'intersection de A et de B est un sous-ensemble de A et un sous-ensemble de B ; c'est aussi un sous-ensemble de tout ensemble contenant A ou B comme sous-ensemble.

(b) $A \cap B = B \cap A$,

i.e. l'intersection de A et de B est égale à l'intersection de B et de A, ce qui tient au fait qu'il revient au même de dire que « x est

élément de A et élément de B » ou de dire que « x est élément de B et élément de A ».

(c) $(A \cap B) \cap C = A \cap (B \cap C)$,

cela veut dire que l'intersection de $A \cap B$ avec C donne le même résultat que l'intersection de A avec $B \cap C$. L'ordre dans lequel on effectue l'opération intersection n'a pas d'influence sur le résultat, ce qui permet de supprimer les parenthèses dans l'écriture :

$$(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C.$$

Cette propriété est à rapprocher de celle du mot « et » dans le langage courant ; en effet, il revient au même de dire : « est élément de A et de B, et est aussi élément de C » que de dire « est élément de A et est élément de B et de C ».

(d) $A \cap A = A$,

i.e. l'intersection de A avec A est A lui-même.

(e) $A \cap \varnothing = \varnothing$,

i.e. l'intersection de A avec l'ensemble vide est l'ensemble vide.

(f) $A \cap B = A \Leftrightarrow A \subset B$,

i.e. l'intersection de A et B est égale à A si, et seulement si, A est un sous-ensemble de B.

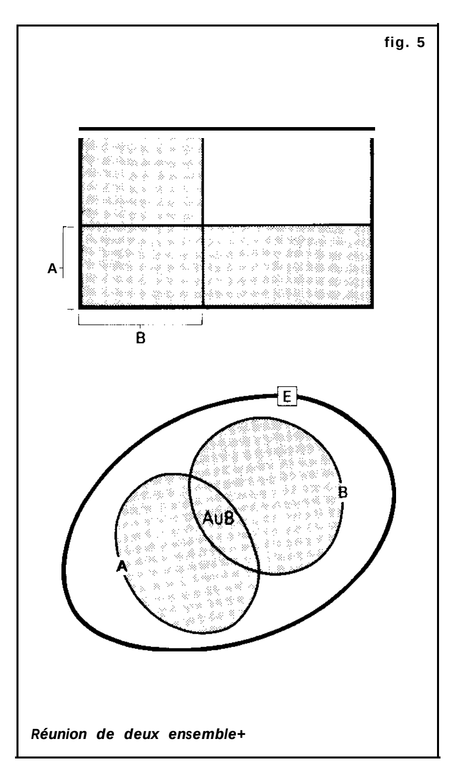
(g) $(D \subset A \text{ et } D \subset B) \Leftrightarrow D \subset A \cap B$,

i.e. tout sous-ensemble commun à A et B est un sous-ensemble de leur intersection et réciproquement.

Réunion de deux sous-ensembles

Si A et B sont deux ensembles, il peut être intéressant, dans certains cas, de « réunir » leurs éléments en un ensemble global. Ainsi, si A désigne l'ensemble des Français possédant une voiture et B l'ensemble des Français possédant un appartement, on peut avoir à considérer l'ensemble des Français qui ont une voiture ou un appartement. Ce nouvel ensemble sera appelé la réunion de A et de B.

La réunion de deux ensembles A et B est un ensemble constitué par les éléments qui appartiennent à A ou à B. On note « $A \cup B$ » ce nouvel ensemble. « $A \cup B$ » se lit « union de A et de B », ou encore « A union B » (fig. 5).



Réunion de deux ensemble+

La définition formelle peut s'écrire :

$$A \cup B = \{x \in E ; x \in A \text{ ou } x \in B\}.$$

Dans cette définition, le mot « ou » joue un rôle très important et toutes les propriétés de la réunion sont des traductions, en langage des ensembles, des propriétés

du mot « ou » au sens non disjonctif dans le langage courant et en logique des propositions.

Voici un exemple de réunion :

$$A = \{ b, l, a, n, c \},$$
$$B = \{ b, l, e, u \},$$
$$A \cup B = \{ b, l, a, e, n, c, u \}.$$

Parmi les propriétés de la réunion, il faut signaler :

(a′) $A \subset A \cup B$ et $B \subset A \cup B$,

i.e. A et B sont des sous-ensembles de la réunion A ∪ B.

(b′) $A \cup B = B \cup A$,

i.e. la réunion de A et de B est égale à la réunion de B et de A, car il revient au même de dire que « x est élément de A ou de B » et de dire que « x est élément de B ou de A ».

(c′) $(A \cup B) \cup C = A \cup (B \cup C)$,

i.e. la réunion de A ∪ B avec C donne le même résultat que la réunion de A avec B ∪ C. L'ordre dans lequel on effectue les réunions n'a pas d'influence sur le résultat ; on peut donc supprimer les parenthèses dans l'écriture :

$$(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C.$$

On pourrait rapprocher cette propriété de celle du mot « ou » dans le langage courant.

(d′) $A \cup A = A$,

i.e. la réunion de A avec A est A.

(e′) $A \cup \emptyset = A$,

i.e. la réunion de A avec l'ensemble vide est A lui-même.

(f′) $A \cup B = B \Leftrightarrow A \subset B$,

i.e. la réunion de A et de B est égale à B si, et seulement si. A est un sous-ensemble de B.

(g′) $(A \subset D \text{ et } B \subset D) \Leftrightarrow A \cup B \subset D$,

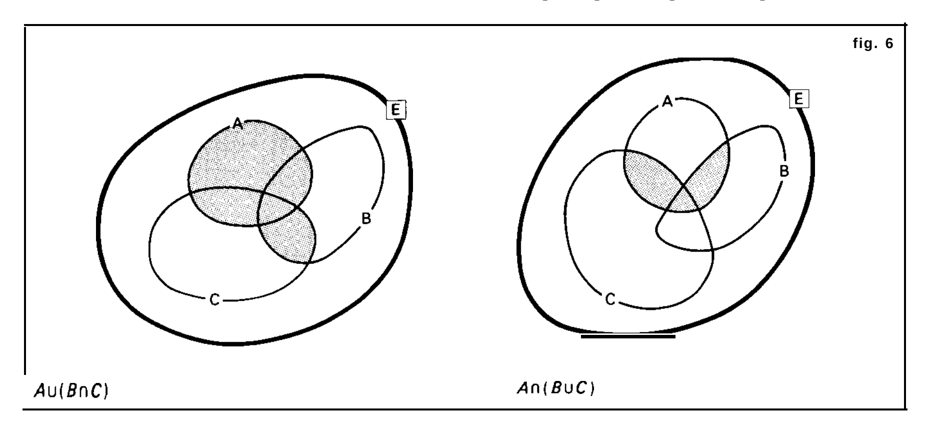
i.e. A et B sont sous-ensembles d'un ensemble D si, et seulement si, la réunion de A et de B est un sous-ensemble de D.

Enfin deux propriétés de distributivité, faisant intervenir la réunion et l'intersection sont à signaler (fig. 6) :

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

Complémentaire

Tous les ensembles considérés ici sont des parties d'un même ensemble E. On **peut** remarquer que chaque fois que l'on définit



une partie de E, on définit automatiquement une autre partie de E, celle qui est constituée par les éléments de E qui ne sont pas dans la première. Ainsi, dans l'ensemble des Français, en définissant l'ensemble des Français inscrits sur les listes électorales, on définit en même temps l'ensemble des Français qui ne sont pas inscrits sur les listes électorales.

On appelle *complémentaire* d'un ensemble A, sous-ensemble de E, l'ensemble constitué par les éléments de E qui ne sont pas dans A (fig. 7). On note « $\bar{A}$ » ou « CA » le complémentaire de A. On peu+ écrire :

$$\bar{A} = \{ x \in E ; x \notin A \}.$$

Quand on parle de complémentaire, il est très important de bien préciser l'ensemble de référence E. Évidemment, le complémentaire de A est un sous-ensemble de E. Entre les parties d'un ensemble E, la complémentarité établit une correspondance un à un, une application bijective (cf. Propriétés des applications in chap. 2) de l'ensemble $\mathcal{P}$ (E) dans lui-même.

Le complémentaire étant défini à partir de la négation, négation de la proposition « appartient à A », ses propriétés sont la traduction, en langage des ensembles, de celles de la négation.

(a″) $$\bar{\bar{A}} = A,$$

i.e. le complémentaire du complémentaire de A est A lui-même.

(b″) $$\bar{\varnothing} = E \text{ et } \bar{E} = \varnothing,$$

i.e. le complémentaire de l'ensemble vide est l'ensemble E, et le complémentaire de l'ensemble E est l'ensemble vide.

(c″) $$A \cap \bar{A} = \varnothing \text{ et } A \cup \bar{A} = E,$$

i.e. l'intersection de A et de son complémentaire est vide, et la réunion de A et de son complémentaire est l'ensemble E. Cela constitue une propriété caractéristique du complémentaire et aurait pu être pris comme définition.

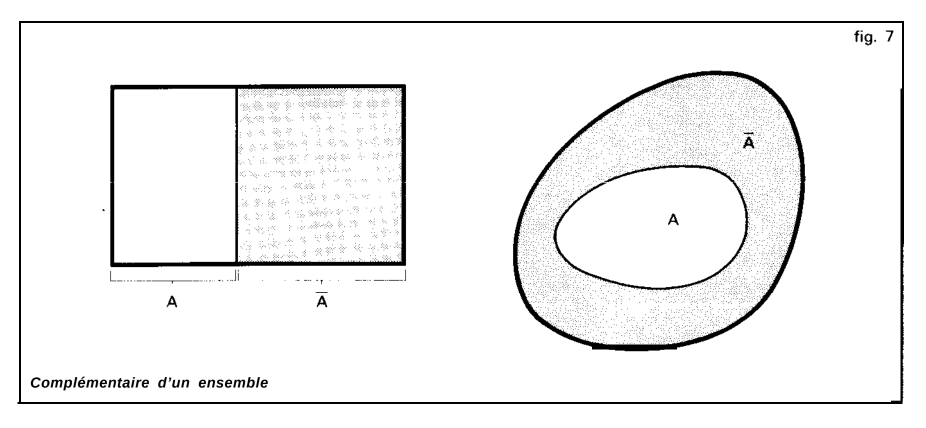
(d″) $$A \subset B \Leftrightarrow \bar{B} \subset \bar{A}.$$

i.e. A est un sc ,-ensemble de B si, et seulement si, le complémentaire de B est un sous-ensemble du complémentaire de A.

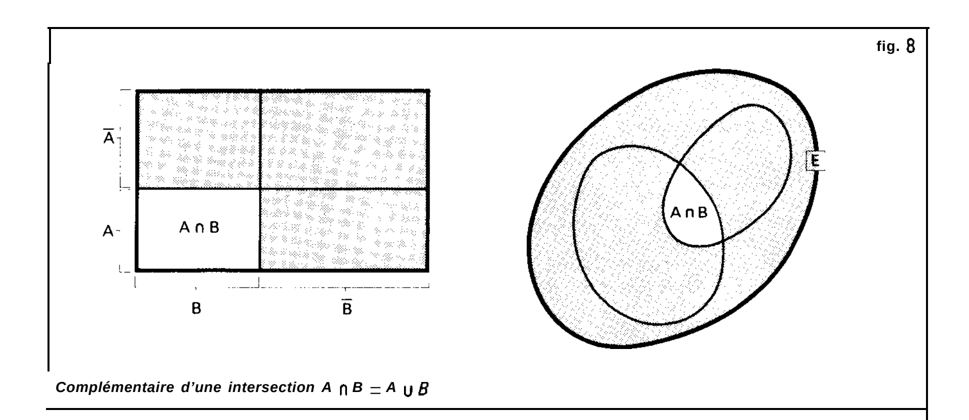
Les lois de De Morgan :

(é') $$\overline{A \cap B} = \bar{A} \cup \bar{B}, \overline{A \cup B} = \bar{A} \cap \bar{B},$$

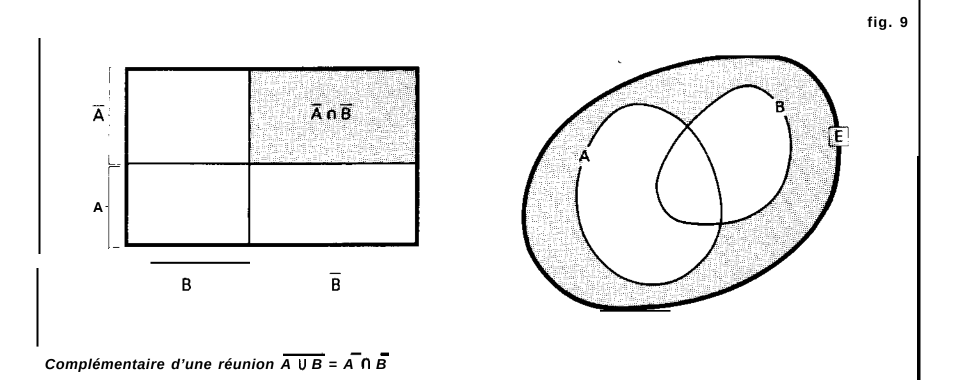
expriment (fig. 8 et 9) que le complémentaire de l'intersection de deux ensembles



Complémentaire d'un ensemble



Complémentaire d'une intersection $\overline{A \cap B} = \overline{A} \cup \overline{B}$



Complémentaire d'une réunion $\overline{A \cup B} = \overline{A} \cap \overline{B}$

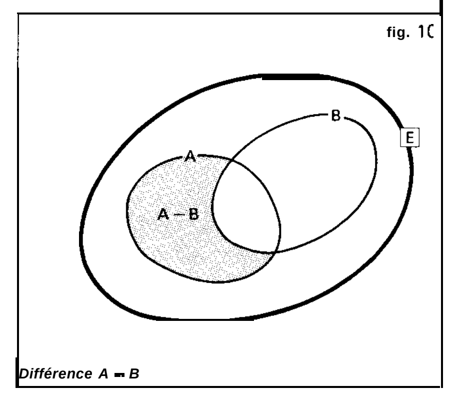
est égal à la réunion des complémentaires de ces deux ensembles, et le complémentaire de la réunion de deux ensembles est égal à l'intersection des complémentaires de ces deux ensembles.

Autres opérations

On peut définir d'autres opérations entre les parties d'un ensemble. Ainsi la *différence* A − B de deux ensembles A et B est l'ensemble constitué par les éléments de A qui ne sont pas éléments de B (fig. 10).

On remarque facilement que :

$$A - B = A \cap \overline{B},$$
$$A - (A - B) = A \cap B,$$
$$A - B = \varnothing \Leftrightarrow A \subset B,$$
$$A - A = \varnothing .$$

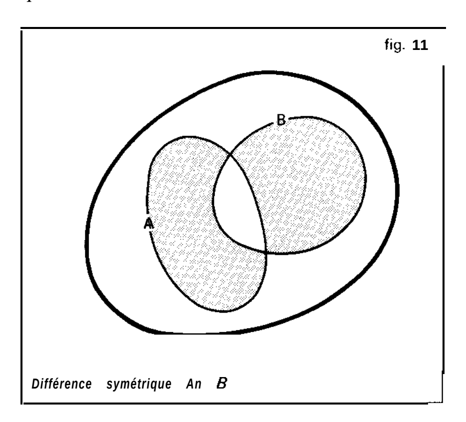


Différence A − B

La *différence symétrique* de deux ensembles A et B est un ensemble constitué par les éléments de A qui ne sont pas dans B et les éléments de B qui ne sont pas dans

A (fig. 11). On le notera A A B, dans ce qui suit.

fig. 11



Différence symétrique An B

On peut vérifier les égalités :

$$AAB = (A-B)U(B-A),$$
$$AAB = (A \cap \bar{B}) \cup (B \cap \bar{A}),$$
$$AAB = BAA,$$
$$(AAB)AC = AA(BAC),$$
$$A \Delta \varnothing = A, \; A \Delta A = \varnothing.$$

Algèbre et onneou de Boole

L'ensemble $\mathcal{P}$ (E) des parties d'un ensemble muni des opérations d'union et d'intersection et de la complémentarité constitue ce qu'on appelle une ***algèbre de Boole.*** En effet, les propriétés suivantes sont vérifiées :

a) Les opérations d'union et d'intersection sont associatives :

$$(A \cup B) \cup C = A \cup (B \cup C),$$
$$(A \cap B) \cap C = A \cap (B \cap C),$$

et commutatives :

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A.$$

b) Il y a un élément neutre pour chacune des deux opérations : pour l'union, $\varnothing$ est élément neutre, car quel que soit l'ensemble A on a : $A \cup \varnothing = A$; et, pour l'intersection, E est élément neutre, car quel que soit A dans $\mathcal{P}$ (E), on a : $A \cap E = A$.

c) Chacune des deux opérations est distributive par rapport à l'autre :

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

d) Le complémentaire $\bar{A}$ de A vérifie les deux propriétés : $A \cup \bar{A} = E$ (élément neutre de l'intersection), et $A \cap \bar{A} = \varnothing$ (élément neutre de la réunion).

Si maintenant on considère l'ensemble T(E) muni des opérations de différence symétrique et d'intersection, il a une structure ***d'anneau de Boole.*** En effet, la différence symétrique donne une structure de groupe commutatif à T(E) ; l'opération intersection donne alors à S(E) une structure d'anneau booléen (cf. **ANNEAUX ET** ALGÈBRES).

Fonctions coroctéristiques

Il est commode de présenter les opérations entre parties d'un ensemble en utilisant les fonctions caractéristiques. Si A est une partie de E, la fonction caractéristique de l'ensemble A est une fonction qui à chaque élément de E associe 1 si cet élément est dans A et 0 si cet élément n'est pas dans A.

Lorsqu'on a la fonction caractéristique d'une partie A de E, il est facile de déterminer celle du complémentaire $\bar{A}$ de A. Il suffit de changer les 0 en 1 et les 1 en 0. Il est aussi facile de voir si un ensemble est inclus dans un autre : on voit que B est inclus dans A si, chaque fois que la fonction caractéristique de B a la valeur 1, celle de A l'a aussi. Les éléments de l'intersection de A et de B sont ceux pour lesquels les fonctions caractéristiques de A et de B prennent ensemble la valeur 1. La fonction caractéristique de $A \cap B$ est donc

facile à construire. Il en est de même pour celle de A ∪ B.

Tout cela peut s'interpréter autrement. En effet, considérons l'ensemble X ayant deux éléments $\{0, 1\}$ et définissons sur cet ensemble deux opérations.

La première opération sera notée + et sera définie par :

$$0 + 0 = 0$$
$$0 + 1 = 1$$
$$1 + 0 = 1$$
$$1 + 1 = 1.$$

La deuxième opération sera notée et sera définie par :

$$0 . 0 = 0$$
$$0 . 1 = 0$$
$$1 . 0 = 0$$
$$1 . 1 = 1.$$

Cet ensemble X muni des deux opérations + et . définies ci-dessus constitue ce qu'on appelle une algèbre de Boole, c'est l'anneau *Z/(2)* des entiers relatifs modulo 2. Il joue un rôle fondamental en logique des propositions. En effet, si on interprète 1 comme le « vrai » et 0 comme le « faux », valeur qu'on peut attribuer aux propositions, les opérations + et . sont alors la disjonction (ou) la conjonction (et) des propositions. Enfin la complémentarité qui est ici l'échange de 0 en 1 et de 1 en 0 s'interprète comme une négation.

Les fonctions caractéristiques sont alors des fonctions qui prennent leur valeur dans l'ensemble X, et si on désigne par f_A la fonction caractéristique de A on a :

$$f_{A \cap B} = f_A . f_B,$$
$$f_{A \cup B} = f_A + f_B,$$
$$f_A + f_{\bar{A}} = f_E = 1,$$
$$f_A . f_{\bar{A}} = f_\emptyset = 0.$$

Remarquons que la formule :

$$f_{A \cup B} = f_A + f_B,$$

ne serait pas vraie si on définissait l'addition des nombres au sens usuel ; on aurait alors $f_{A \cup B} = f_A + f_B - f_A f_B$.

Partition d'un ensemble

Lorsqu'on a un ensemble d'objets ayant chacun une couleur bien déterminée, on peut être amené à les classer suivant leur couleur. On effectue ainsi une classification des objets, chaque objet étant dans une classe et une seule. On a de la sorte une image de ce que le mathématicien appelle une partition.

Une partition d'un ensemble est un ensemble de parties non vides de cet ensemble tel que deux parties distinctes n'aient pas d'élément commun et que chaque élément de l'ensemble soit dans une de ces parties. C'est en quelque sorte l'ensemble des classes d'une classification.

Ainsi, dans l'ensemble des entiers naturels N = { 1. 2, 3, 4, }, on peut constituer une partition en trois classes en plaçant dans la première classe les nombres divisibles par 3, dans la deuxième ceux dont le reste de la division par 3 est 1, dans la troisième ceux dont le reste est 2. Ainsi chaque nombre est rangé dans une classe et aucun ne se trouve dans deux classes en même temps. Les classes sont alors : { 0, 3, 6, 9, 12, 15, ... }, { 1, 4, 7, 10, ... } et { 2, 5, 8. 11, }.

On trouve de nombreux exemples de partitions dans les problèmes de rangement, les dépouillements de questionnaires, les classifications, etc.

Quelques applications

Circuits électriques

L'algèbre de Boole étudiée ci-dessus. ou algèbre des parties d'un ensemble, peut être appliquée à l'étude des circuits électriques. Ainsi, dans un circuit relié aux

deux pôles d'une pile par exemple, si on place un interrupteur à deux positions, deux cas peuvent se présenter : ou bien l'interrupteur est fermé et le courant passe, on peut noter cela par 1 ; ou bien l'interrupteur est ouvert et le courant ne passe pas, on peut noter ceci par 0. On peut alors définir des opérations sur les interrupteurs analogues à celle que nous avions dans l'ensemble X = { 0.1 } ci-dessus.

Deux interrupteurs I et J disposés en série ne laissent passer le courant que s'ils sont fermés tous les deux. Notons I . J cette combinaison en série des deux interrupteurs 1 et J ; l'état de I J est alors défini par la table :

1	J	I · J
1	1	1
1	0	0
0	1	0
0	0	0

De même, si deux interrupteurs sont placés en parallèle, le courant passe dès que l'un *ou* l'autre des interrupteurs est fermé. Si on note I + J l'état du circuit, I + J est défini à partir de 1 et de J comme l'indique le tableau suivant :

I	J	I + J
1	1	1
1	0	1
0	1	1
0	0	0

Enfin on peut retrouver la négation par un circuit complémentaire $\bar{I}$, le tableau donnant l'état de chacun des circuits I et $\bar{I}$:

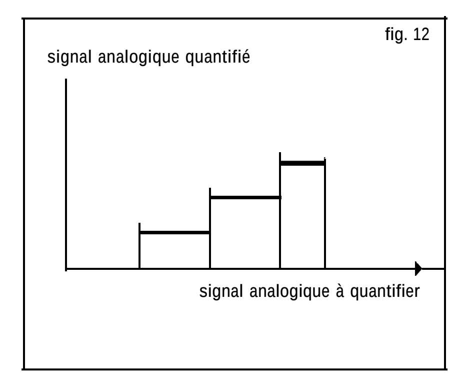
1	$\bar{I}$
1	0
0	1

Cela peut permettre d'étudier les circuits électriques, mais, réciproquement, c'est ce qui a permis l'utilisation des circuits électriques dans les machines à calculer électroniques.

Informotion analogique

L'information analogique décrite mathématiquement par des fonctions continues (signal électrique issu d'un microphone) doit être quantifiée et codée pour être traitée par les procédés de l'électronique moderne utilisant largement les microprocesseurs et le traitement numérique de l'information.

Le principe de la quantification est le suivant. Au signal analogique est associé un nombre d'états possibles selon le graphe de la figure 12



Le nombre d'états possibles de quantification est en général élevé, 2^{12} par exemple. Le traitement numérique impose d'associer à chaque état de quantification une suite de valeurs booléennes (12 variables booléennes peuvent être utilisées) ; l'information est alors représentee par une suite d'états 0 ou 1.

Une application connue de tous est le disque compact ; celui-ci permet de stocker les suites binaires sous forme de « trous » situés sur le sillon du disque. Le nombre de 0 et de 1 stockés sur le disque compact est très élevé : 3 000 mégaéléments binaires !

L'évolution est loin d'être terminée : les images sont stockables en utilisant le même procédé. et l'on attend beaucoup de la logique floue (fondements mathématiques dérivés de l'algèbre de Boole, mais différents) et des réseaux de neurones.

ANDRÉ ROUMANET

2. Relations

Produit cartésien

te couple

Soit E et F deux ensembles, Pour $x \in E$ et $y \in F$, on introduit un nouvel objet mathématique, le *couple* de premier terme x et de second terme y, défini par le symbole :

$$(x, y),$$

avec la convention que :

$$(x, y) = (x', y') \Leftrightarrow x = x' \text{ et } y = y'.$$

On appelle *produit cartésien* de deux ensembles E et F, noté E X F, l'ensemble des couples ayant pour premier terme un élément de E et pour second terme un élément de F. Par exemple, si $E = \{a, b, c\}$ et $F = \{A, B\}$ sont des ensembles à trois et deux éléments respectivement, l'ensemble E X F a six éléments qui sont :

$$(a, A), (a, B), (b, A), (b, B), (c, A), (c, B),$$

et l'ensemble E X E a neuf éléments qui sont :

$$(a,a), (a, b), (a,c), (b, a), (b,b), (b, c), (c,a), (c, b), (c,c) ;$$

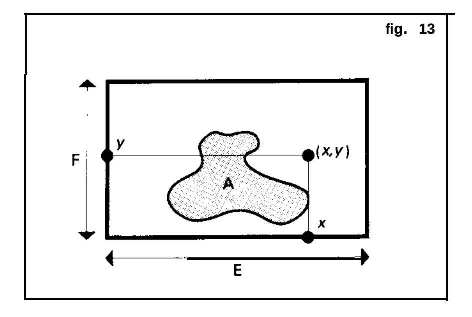
plus généralement, si E et F sont des ensembles finis contenant m et n éléments, le produit cartésien E X F est fini et contient mn éléments.

Justifions la terminologie de cartésien. Le choix de deux axes de coordonnées dans le plan de la géométrie élémentaire permet d'identifier l'ensemble des points du plan à l'ensemble R X R = $\mathbf{R}^2$ des couples de nombres réels, au point M correspondant le couple ayant pour premier terme l'abscisse de M et pour second terme son ordonnée ; c'est le principe de la géométrie analytique de Descartes, chez qui apparaît pour la première fois la notion mathématique de couple.

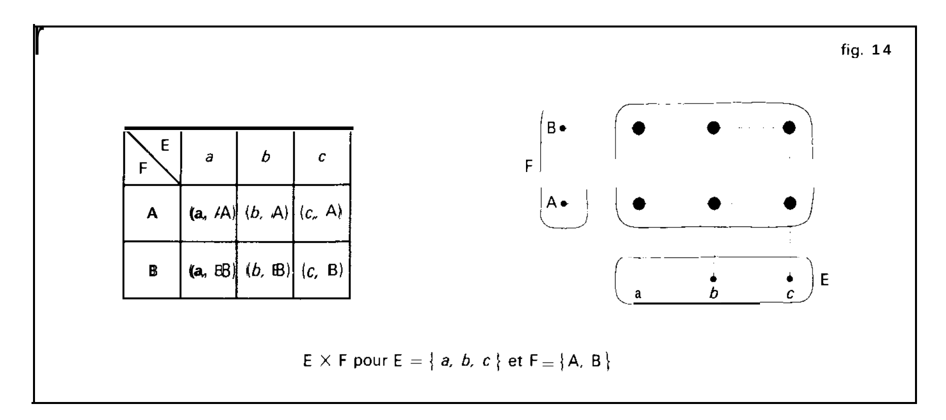
De nos jours, on définit souvent ainsi le plan de la géométrie élémentaire ; dans ce qui suit, cette identification sera toujours faite.

Représentations graphiques

On représente souvent (représentation dite cartésienne) un ensemble produit E X F par l'ensemble des points d'un rectangle (surtout ne pas confondre avec les diagrammes de Carroll !), les ensembles E et F étant représentés par deux côtés perpendiculaires de ce rectangle (fig. 13) ; un sous-ensemble A de E X F est alors représenté par un sous-ensemble de ce rectangle.



Dans le cas d'ensembles finis, on peut faire le tableau donnant les éléments de l'ensemble produit ou utiliser une représentation par des points du plan (cf. fig. 14 pour l'exemple ci-dessus). Pour représen-



E × F pour E = {a, b, c} et F = {A, B}

ter les sous-ensembles, on peut indiquer leurs éléments sur la représentation, mais on peut aussi utiliser la représentation sagittale, dont voici le principe : on représente le couple (x, y) par deux points (appelés x et y) réunis par une flèche allant de x vers y ; dans le cas particulier d'un couple (x, x), on dessine une boucle fermée allant de x à x. Sur la figure 15, on donne les représentations cartésienne et sagittale du sous-ensemble :

$$X = \{(a, B), (b, A), (c, A), (c, B)\}$$

de l'ensemble E X F vu ci-dessus.

Relations binaires

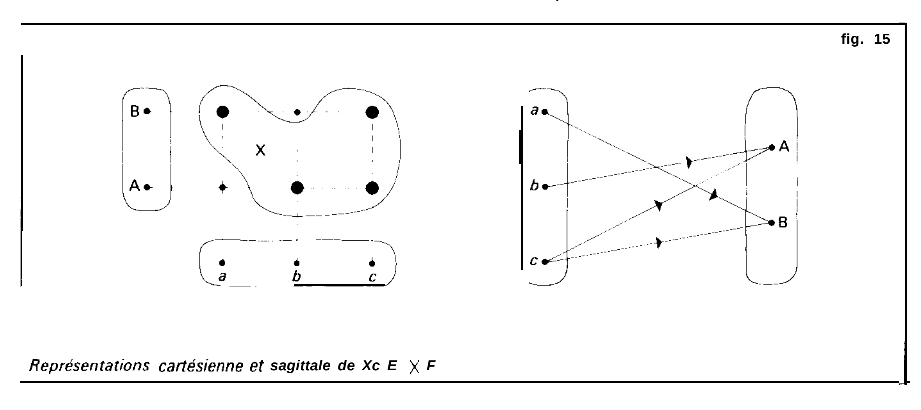
Soit E et F des ensembles. Une *relation* de *source* E et *but* F est une propriété sur l'ensemble produit E X F, c'est-à-dire une propriété des couples $(x, y), x \in E$ et $y \in F$. Ainsi une relation définit un sous-ensemble de E X F, appelé son *graphe*, formé des couples pour lesquels la relation est vraie (cf. chap. 1). Réciproquement, tout sous-ensemble A C E X F définit une relation de source E et de but F, à savoir la propriété :

$$(x, y) \in A$$

du couple (x, y). Lorsque F = E, on dit qu'on a une relation sur E. Si une relation $\mathcal{R}$ est vraie pour le couple (x, y), on écrira souvent $x \mathcal{R} y$.

Exemples

(1) Sur un ensemble E, la relation d'égalité « $x = y$ » a pour graphe l'ensemble des



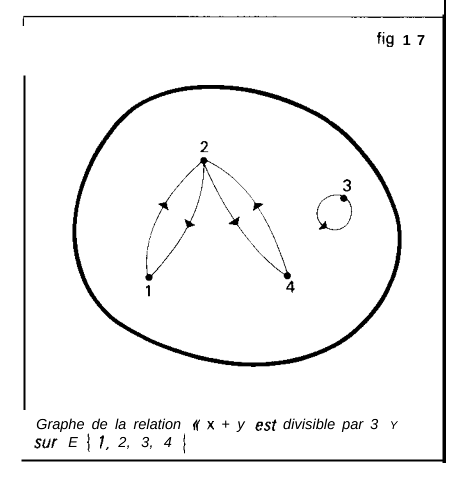
Représentations cartésienne et sagittale de X c E × F

fig. 16

couples (x, x), $x \in E$; cet ensemble est appelé la *diagonale* de l'ensemble $E \times E$.
(2) Prenons pour E l'ensemble des quatre voyelles $\{a, e, i, o\}$ et pour F l'ensemble des quatre premiers chiffres $\{1, 2, 3, 4\}$; la figure 16 indique les représentations cartésienne et sagittale de la relation « la voyelle x figure dans l'écriture en langue française du chiffre y ». Par exemple (a, 4) appartient au graphe (car a est dans quatre), mais (a, 3) ne lui appartient pas.
(3) Prenons $E = F = \{1, 2, 3, 4\}$ et la relation sur E « $x + y$ est divisible par 3 ». Le graphe contient cinq éléments, qui sont :

$$(1, 2), (2, 1), (2, 4), (4, 2), (3, 3);$$

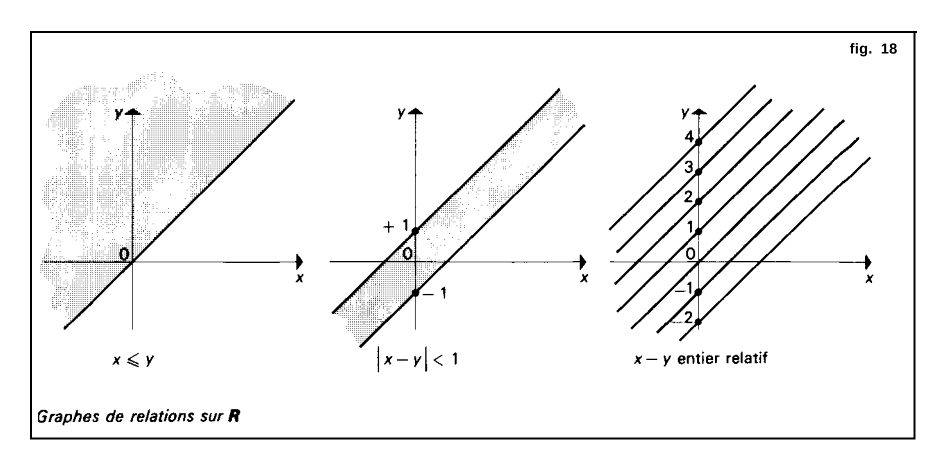
la figure 17 donne la représentation sagittale de ce graphe : remarquer la boucle pour représenter (3, 3).
(4) Sur l'ensemble R des nombres réels, soit la relation « $x \leqslant y$ ». Si on identifie, comme indiqué ci-dessus, l'ensemble des couples de nombres réels aux points du plan, le graphe est constitué des points de la première bissectrice $y = x$ et des points situés au-dessus de cette droite (fig. 18).
(5) Toujours sur R, considérons maintenant la relation « $|x - y| < 1$ ». Son graphe est ici la bande du plan comprise entre les deux droites $x - y = 1$ et $x - y = -1$ (fig. 18).
(6) Encore sur R, soit la relation « $x - y$ est un entier relatif » ; le graphe de cette relation est le sous-ensemble du plan formé des droites $y = x + n$, lorsque n parcourt l'ensemble des entiers relatifs (fig. 18).



Graphe de la relation « $x + y$ est divisible par 3 » sur $E \{1, 2, 3, 4\}$

Propriétés des relations

On se limite maintenant à des relations sur un ensemble E et on se propose de dégager un certain nombre de propriétés de ces relations.



Graphes de relations sur ***R***

Une relation sur E est dite *réflexive* si, pour tout élément $x \in$ E, la relation est vraie pour le couple (x, x) ; cela revient donc à dire que tous les couples (x, x) appartiennent au graphe de la relation, ou encore que ce graphe contient la diagonale de l'ensemble E X E. Les relations données ci-dessus sont réflexives dans les exemples (1), (4), (5), (6), non réflexives dans l'exemple (3).

Une relation sur E est dite *symétrique* si, toutes les fois que la relation est vraie pour un couple (x, y), elle est aussi vraie pour le couple (y, x). Ainsi les relations des exemples (1), (3), (5), (6) sont symétriques, celle de l'exemple (4) ne l'est pas. À l'opposé, une relation est dite *antisymétrique* si on a nécessairement $x = y$ quand la relation est vraie à la fois pour le couple (x, y) et pour le couple (y, x) ; c'est le cas de (4).

Une relation sur E est dite *transitive* si. toutes les fois que la relation est vraie pour des couples (x, y) et (y, z), elle est aussi vraie pour le couple (x, z). Ainsi, les relations des exemples (1), (4), (6) sont transitives, celles des exemples (3) et (5) ne le sont pas : ainsi, dans l'exemple (3), la relation est vraie pour les couples (1, 2) et (2, 4), mais ne l'est pas pour le couple (1, 4).

Avec ces notions, on peut désormais distinguer différents types de relations. Les *relations d'ordre* sont les relations qui, comme dans l'exemple (4), sont réflexives, antisymétriques et transitives ; ces relations sont très importantes, mais nous ne les étudierons pas ici, et nous renvoyons à l'article ensembles **ORDONNÉS**. Les relations d'équivalence sont celles qui, comme dans l'exemple (6), sont réflexives, symétriques et transitives (cf. *infra*, *Relutions d'équivalence*).

Applications d'un ensemble dans un autre

Définitions

On revient à la situation générale de deux ensembles **E** et F et d'une relation de source E et de but F.

Nous dirons qu'une relation de source E et de but F est une *application* de E dans F si, pour tout élément $x \in$ E, il existe un *unique* élément $y \in$ F tel que la relation soit vraie pour (x, y). Ainsi à tout élément $x \in$ E correspond un unique élément $y \in$ F appelé *l'image* de x par

l'application. Si l'application est désignée par f, on notera $y = f(x)$ l'image de x ; on écrira :

$$f : \mathrm{E} \to \mathrm{F}$$

pour exprimer que f est une application de E dans F, et :

$$x \mapsto y = f(x)$$

pour exprimer que y est l'image de x par f. Si E et F sont des ensembles finis. la flèche de la représentation sagittale représente le passage de x à son image $f(x)$.

La relation d'égalité sur un ensemble E donne un exemple trivial d'application ; l'image de tout élément $x \in$ E est cet élément x lui-même. On obtient ainsi ce qu'on appelle l'*application identique* de E dans lui-même, notée 1,.

On introduit aussi souvent, non sans quelques confusions, la notion de relation fonctionnelle, ou de fonction. Une relation de source E et de but F est une *fonction* si, pour tout élément $x \in$ E, il existe au *plus* un élément y $\in$ F pour lequel la relation est vraie. On appelle alors *ensemble de définition* d'une telle fonction le sous-ensemble E' de E formé des éléments $x \in$ E pour lesquels il existe effectivement un tel y ; si on fait correspondre à tout élément x $\in$ E' cet élément y $\in$ F, on définit une application de E' dans F, mais il est fondamental, si E' $\neq$ E, de ne pas confondre la fonction initiale de l'application que nous venons de lui associer, car la première relation est de source E et la seconde de source E'. Par exemple, considérons sur R la relation :

$$y = \frac{x + 1}{x - 1} ;$$

c'est une fonction dont le domaine de définition est l'ensemble E' des nombres réels différents de 1 et elle définit une application f : E' $\to$ R telle que :

$$x \mapsto \frac{x + 1}{x - 1} .$$

Si f : E $\to$ F est une application de E dans F, on appelle graphe de f son graphe au sens généra1 défini ci-dessus pour une relation. C'est le sous-ensemble de E X F formé des couples $(x, f(x))$, $x \in$ E ; ainsi, si G est ce graphe :

$$(x, y) \in \mathbf{G} \Leftrightarrow y = f(x).$$

Dans le cas où E C R et F = R, le graphe est le sous-ensemble de $\mathbf{R}^2$ formé des couples (x, y) tels que $x \in$ E et $y = f(x)$; c'est le graphe, au sens usuel, de f.

Composition des applications

Soit f : E $\to$ F et g : F $\to$ G deux applications ; remarquons ce qui est essentiel ici, que l'ensemble source de g est le même que l'ensemble but de f. Considérons la propriété suivante des couples $(x, z) \in$ E X G : il existe un élément $y \in$ F tel que $y = f(x)$ et $z = g(y)$; cette relation sur E X G est en fait une application de E dans G. En effet, pour x $\in$ E donné, l'élément y = $f(x)$ est parfaitement déterminé, et la relation exprime que $z = g(f(x))$. Ainsi, à tout x $\in$ E cette application, appelée application composée de f et g et notée g $\circ f$ (dans cet ordre) fait correspondre à x l'unique élément z image par g de l'élément $y = f(x)$. Pratiquement, pour trouver l'image par g $\circ$ f d'un élément x $\in$ E, on cherche donc d'abord l'image y de x par f, puis l'image par g de cet élément y.

Soit par exemple les ensembles $E = \{a, b, c\}$, $F = \{\alpha, \beta, \gamma, \delta\}$ et $G = \{1, 2, 3\}$; considérons $f : E \rightarrow F$ et $g : F \rightarrow G$ telles que :

$$f(a) = \beta,\ f(b) = \alpha,\ f(c) = \delta$$

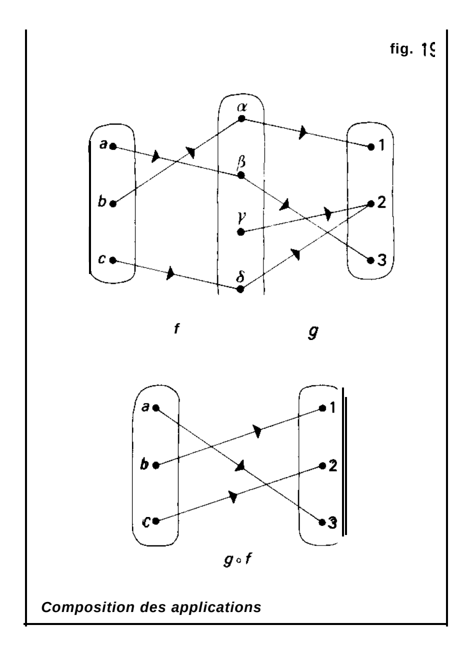
et

$$g(a) = 1,\ g(\beta) = 3,\ g(\gamma) = 2,\ g(\delta) = 2 ;$$

l'application $g \circ f$ est telle que :

$$g \circ f(a) = 3,\ g \circ f(b) = 1,\ g \circ f(c) = 2$$

(la figure 19 représente les diagrammes sagittaux des graphes de ces applications)



Composition des applications

Avec E = F = G = R, voici un autre exemple. Soit $f : R \rightarrow R$ et $g : R \rightarrow R$ définies par $f(x) = x^2$ et $g(x) = 2x + 1$; ici on peut composer f et g, mais aussi g et f. Ces deux applications de R dans R sont définies par $g \circ f(x) = 2x^2 + 1$ et $f \circ g(x) = (2x + 1)^2$; il s'agit d'applications distinctes.

Propriétés des applications

Une application $f : E \rightarrow F$ est dite *surjective* si, pour tout élément $y \in F$, il existe au moins *un* élément $x \in E$ ayant pour image y, c'est-à-dire tel que $y = f(x)$; on dit souvent que f est une application de E sur F, ou est une *surjection.* Ainsi, dans le premier exemple donné ci-dessus (cf. « Composition des applications »), l'application g est surjective ainsi que l'application $g \circ f$; l'application f, au contraire, ne l'est pas, car il n'existe aucun élément de E ayant pour image y.

Une application $f : E \rightarrow F$ est dite *injective* si, pour tout élément $y \in F$, il existe *au plus* un élément $x \in E$ ayant pour image y; cela revient à dire que deux éléments distincts de E ont des images qui sont des éléments distincts de F. On dit souvent que f est une *injection* de E dans F. Ainsi, toujours en reprenant le premier exemple donné dans « Composition des applications », les applications f et $g \circ f$ sont injectives, tandis que g ne l'est pas (en effet, γ et δ ont même image par l'application g).

Attirons l'attention ici sur l'importance, pour définir une application, de préciser les ensembles source et but. Ainsi, désignant par $\mathbf{R}^+$ l'ensemble des nombres réels positifs, considérons les trois applications :

$f : R \rightarrow R$, définie par $f(x) = x^2$,
$g : R \rightarrow R^+$, définie par $g(x) = x^2$,
$h : R^+ \rightarrow R$, définie par $h(x) = x^2$,

ces trois applications sont *distinctes* et possèdent des propriétés différentes : f n'est ni injective ni surjective, g est surjective, mais n'est pas injective, et h est injective, mais n'est pas surjective.

Bijections

Une application $f : E \rightarrow F$ qui est à la fois injective et surjective est dite *bijective ;* on

dit aussi que f est une *bijection*. Les bijections jouent un rôle très important en théorie des ensembles (construction des cardinaux par exemple) ; du point de vue de la théorie des ensembles proprement dite, on peut « identifier » des ensembles E et F tels qu'il existe une bijection de E sur F (ce que avons fait plus haut pour le plan et R X R).

Si $f : E \longrightarrow F$ est une bijection, pour tout élément $y \in F$ il existe un unique élément $x \in E$ tel que $y = f(x)$; la propriété du couple $(u, v) \in F \times E$:

$$u = f(v)$$

est une relation sur F X E qui est en fait une bijection de F sur E. En effet, puisque f est une bijection, pour tout $u \in F$ il existe un unique élément $v \in E$ ayant pour image u ; la bijection ainsi définie est appelée la bijection réciproque de la bijection f, notée f^{-1}. C'est donc une application de F dans E, et on a la caractérisation :

$$u = f(v),\ v \in E \Leftrightarrow v = f^{-1}(u),\ u \in F.$$

On disait autrefois que f ou f^{-1} établissaient une correspondance biunivoque entre E et F ; en effet, les formules précédentes associent deux à deux les éléments de E et ceux de F. Remarquons que l'on peut composer $f^{-1} \circ f$ et $f \circ f^{-1}$; on obtient ainsi respectivement les applications identiques (c'est-à-dire la relation d'égalité) dans E et F respectivement.

Dans le premier exemple du paragraphe « Composition des applications », l'application $g \circ f$ est une bijection de $E = \{a, b, c\}$ sur $G = (1, 2, 3\}$; l'application réciproque serait :

$$1 \mapsto b,\ 2 \mapsto c,\ 3 \mapsto a.$$

Relations d'équivalence

On appelle *relation d'équivalence* sur un ensemble E une relation sur E qui est réflexive, symétrique et transitive. Si une relation d'équivalence donnée est vraie pour un couple (x, y), on dit que ces éléments sont équivalents (modulo la relation considérée) et on note $x \sim y$.

Exemples

(1) Sur l'ensemble $E = \{1, 2, 3, 4\}$, considérons la relation « x y est un multiple de 3 » (dans les entiers relatifs). On vérifie facilement que c'est une relation d'équivalence ; la figure 20 représente le diagramme sagittal de cette relation.

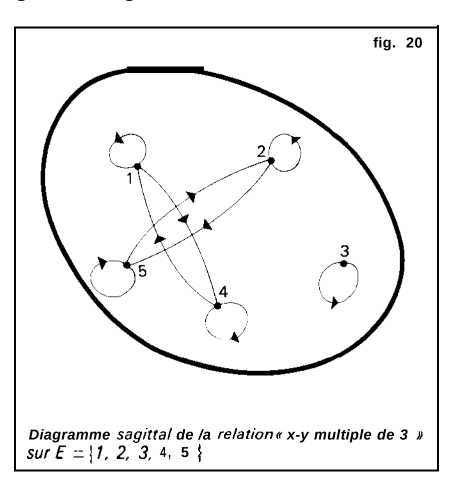


fig. 20

Diagramme sagittal de la relation « x-y multiple de 3 » sur E = {1, 2, 3, 4, 5}

(2) Voici un exemple plus général (en fait, toutes les relations d'équivalence peuvent être obtenues ainsi). Soit $f : E \longrightarrow F$ une application d'un ensemble E dans un ensemble F. La relation sur E :

$$x \mathcal{R} y \Leftrightarrow f(x) = f(y)$$

est une relation d'équivalence dite *associée à l'application f*.

(3) La donnée d'une « fraction » p/q équivaut à la donnée de son numérateur p,

qui est un entier relatif quelconque, et de son dénominateur, qui est un entier relatif non nul, c'est-à-dire qu'elle équivaut à la donnée du couple $(p, q), p \in Z$ et $q \in Z^* = Z - \{0\}$. Dans ce qui suit, nous identifierons donc l'ensemble des fractions à l'ensemble produit $Z \times Z^*$. Sur cet ensemble, la relation :

$$(p, q) \sim (p', q') \Leftrightarrow pq' = p'q$$

(le produit des « extrêmes » est égal au produit des « moyens ») est une relation d'équivalence.

(4) Sur l'ensemble des droites du plan, la relation $D // D'$, qui exprime que les droites D et D' sont parallèles ou confondues, est une relation d'équivalence.

(5) Sur l'ensemble des « vecteurs » du plan, c'est-à-dire des couples de points du plan (origine, extrémité), la relation *d'équipollence* « $\overrightarrow{AB} \sim \overrightarrow{CD}$ si les segments AD et BC ont même milieu » est une relation d'équivalence.

Ensemble quotient

Soit E un ensemble muni d'une relation d'équivalence. Pour tout élément $x \in E$, on appelle *classe d'équivalence de* x l'ensemble, noté C, ou $\dot{x}$, des éléments de E qui sont équivalents à x ; c'est le sous-ensemble de E :

$$C_x = \{y \in E ; y \sim x\}$$

qui est toujours non vide, car il contient x (réflexivité).

Remarquons que tous les éléments d'une même classe d'équivalence sont équivalents entre eux et que deux éléments équivalents ont des classes égales (symétrie et transitivité de la relation d'équivalence). Il en résulte que si deux éléments x et y ne sont pas équivalents, leurs classes d'équivalence sont disjointes, c'est-à-dire n'ont pas d'élément commun.

Puisque les classes d'équivalence sont des sous-ensembles de E, l'ensemble de ces classes d'équivalence est un sous-ensemble de l'ensemble $\mathcal{P}$ (E) des parties de E ; on appelle ensemble *quotient* de E par la relation d'équivalence considérée ce sous-ensemble de $\mathcal{P}$ (E), qui admet pour éléments les classes d'équivalence. Remarquons que l'ensemble quotient est une *partition* de l'ensemble E : les classes d'équivalence sont toutes non vides, et tout élément de E appartient à une classe d'équivalence et une seule, la sienne (et celle de tous les éléments qui lui sont équivalents). Réciproquement, si $\mathcal{F} \subset \mathcal{P}(E)$ est une partition, on peut lui associer la relation d'équivalence : « x et y appartiennent au même sous-ensemble de la partition » ; bien entendu, on retrouve comme ensemble quotient par cette relation d'équivalence la partition initiale, la classe d'un élément étant ici l'unique ensemble de la partition auquel il appartient.

Reprenons, avec ces nouvelles notions, les exemples que l'on vient de donner. Dans l'exemple (1), on a :

$$C_1 = C_4 = \{1, 4\}, C_2 = C_5 = \{2, 5\}, C_3 = \{3\} ;$$

ainsi l'ensemble quotient contient trois éléments, qui sont $\{1, 4\}$, $\{2, 5\}$ et $\{3\}$. Les exemples (3) à (5) montrent comment une relation d'équivalence permet de *définir* de nouveaux objets mathématiques. Dans (3), toutes les fractions équivalentes entre elles correspondent au même nombre rationnel ; *par définition*, on appellera nombre rationnel tout élément de l'ensemble quotient : on a ici une construction mathématique rigoureuse des nombres rationnels à partir des entiers relatifs. Les exemples (4) et (5) permettraient de définir mathématiquement les notions de direction de droite et de vecteur libre respectivement.

Représentants

Si C est un élément de l'ensemble quotient, on appelle *représentant* de cette classe tout élément $x \in C$, c'est-à-dire tout élément de E tel que $C_x = C$. Comme il est souvent plus facile de manipuler les éléments de E que les éléments de l'ensemble quotient, il est intéressant de trouver des représentants de **toutes** les classes d'équivalence, « sans omission ni répétition ». Un sous-ensemble S de E est appelé un *système complet de représentant.~* si tout élément de E est équivalent à un élément de S et à un seul. Ainsi, dans l'exemple (1) ci-dessus, on peut prendre $S = \{1, 2, 3\}$.

Examinons l'exemple (2). Une fraction p/q sera dite *irréductible* si $q > 0$ et si p et q sont premiers entre eux. Les propriétés élémentaires des entiers montrent que chaque classe d'équivalence contient une fraction irréductible et une seule, toutes les fractions équivalentes étant de la forme kp/kq, $k \in Z^*$; ainsi, les fractions irréductibles forment un système complet de représentants des nombres rationnels.

Dans l'exemple (5), fixons un point 0 du plan ; dans chaque vecteur libre, il existe un vecteur et un seul d'origine 0 (ce qu'on exprime dans le langage usuel en disant que le choix de son origine détermine complètement un vecteur libre) ; ainsi les vecteurs d'origine 0 forment un système complet de représentants des vecteurs libres du plan.

JEAN-LUC VERLEY

Bibliographie

M. AUMIAUX, *Logique binaire : fonctions logique.~ et arithmétique binaire,* Masson, 2' éd. 1992 / M. BARBUT. *Mathématiques des science.~ humaines*, 2 vol., P.U.F., 4' éd. 1975-1976 / S. BARUK, *Dictionnuire des mathématiques*, Le Seuil, 1992 / N. BOURBAKI. *Éléments d'histoire des mathématiques*, Masson, 1984 / A. BOWIER, *Lu Théorie des ensembles,* coll. Que sais-je ?, P.U.F., 3' éd. 1982 / L. CARROLL, *Logique sans peine*, Hermann, 3e éd. 1972 / L. CHAMBADAL, *Dictionnuire de mathématiques,* Hachette, Paris, 1981 / C. JEULIN, R. PROTEAU, D. SPERANDIO et al., *Ensembles, relations,* Vuibert, 1981 / J. PERMINGEAT & D. GLAUDE, *Algèbre de Boole : théorie, méthodes de calcul, applications*, Masson, 1991 / J. PICHON, *Théorie des ensembles, logique, les entiers*, Marketing, 1989.

ENSEMBLES CONVEXES

→ CONVEXITÉ - Ensembles convexes

ENSEMBLES ORDONNÉS

→ ORDONNÉS ENSEMBLES

ÉQUATIONS ALGÉBRIQUES

Dès la plus haute antiquité, on rencontre, à l'occasion de problèmes concrets, des exemples de résolution d'équations du premier et du second degré, et, jusqu'au début du XIXe siècle, l'étude des équations constitue l'unique préoccupation des algébristes.

Le développement de la théorie est étroitement lié aux extensions successives de la notion de *nombre* : introduction des nombres négatifs, des nombres irrationnels, tandis que les formules de résolution de l'équation du troisième degré allaient conduire les algébristes italiens du XVIe siècle à raisonner sur les nombres imaginaires (cf. nombres COMPLEXES).

Par analogie avec le cas des équations de degré inférieur ou égal à 4, les algébristes pensèrent que toute solution d'une équation pouvait s'exprimer par des radicaux portant sur les coefficients de l'équation. Par un hasard de l'histoire des sciences, les tentatives pour établir cette conjecture, pourtant mathématiquement saugrenue, allaient conduire à dégager les premières structures abstraites et être à l'origine de l'algèbre moderne (cf. ALGÈBRE).



1. Équations affines

On étudiera en premier lieu le développement historique des systèmes d'équations affines.

Premier exemple

Problème 69 du *Papyrus Rhind* (Égypte), vers 1700 avant notre ère : « Trois boisseaux et demi de farine sont transformés en 80 pains. Dis-moi combien chaque pain contient de farine et quelle est leur force. »

Rappelons que le boisseau *(hequt)* mesure environ 4,5 litres. Il est divisé en 1/2, 1/4, 1/8, 1/16, 1/32, 1/64 de boisseau et contient 320 « ros » (ou parties). La « force » d'un pain est la quantité de pains que peut fournir un boisseau de farine. Si x est cette force et si y est la quantité de farine contenue dans un pain, x et y sont avec les mêmes unités inverses l'un de l'autre. Le texte donne pour la force :

$$80 : 3\frac{1}{2} = \frac{2}{3}\,\frac{1}{7}\,\frac{1}{21};$$

et, pour la quantité de farine contenue dans un pain, 3 boisseaux et demi divisés par 80, ou 1 120 ros divisés par 80, donnent 1/32 de boisseau et 4 ros.

C'est un problème très élémentaire du type $ax = b$ ou $a = by$. Toute la difficulté provient, au point de vue concret, du choix des unités de mesure et de leurs subdivisions, et, au point de vue abstrait, du calcul égyptien des fractions. Dans ce calcul, la notion de fraction générale n'est pas encore dégagée, ou, en langage actuel, l'ensemble $\mathbf{Q}^+$ n'est pas mis en évidence. À part la fraction 2/3, l'Égyptien ne calcule que par quantièmes ou fractions de numérateur 1. Ces errements se prolongeront très longtemps dans les littératures mathématiques grecque (collection héronienne), byzantine et occidentale.

Deuxième exemple

On peut trouver en Égypte des problèmes plus savants que le précédent, qui se ramènent au même type d'équation. Prenons cependant, dans la mathématique babylonienne, un deuxième exemple à peu près contemporain du précédent (E. Bruins et M. Rutten, *Textes mathématiques de Suse*) : « Un quart de la largeur, ajoute à la longueur : 7 mains... à 10... 10 c'est la somme. Largeur ? » En désignant la longueur par x, la largeur par y, on obtient le système $x + y/4 = 7$; $x + y = 10$. Voici la solution donnée dans la tablette : « Porte 7 à 4 du « quart » : 28 tu trouves ; tu soustrais 10 de 28 : 18 tu trouves. Dénoue l'inverse de 3 : 20 tu trouves ; porte 20 à 18 : 6 tu trouves : 6 la longueur ; tu soustrais 6 de 10 : 4, la largeur.. »

Ce système de deux équations à deux inconnues est résolu suivant un procédé encore utilisé dans notre enseignement

élémentaire. La numération utilisée est à base 60. La division est remplacée par la multiplication par l'inverse du diviseur.

Troisième exemple

La littérature chinoise offre, dans le même ordre d'idées, des exemples ultérieurs, parmi lesquels le suivant, extrait de Neuf *Chapitres sur l'art du calcul*, ouvrage qui se situe dans les deux derniers siècles avant notre ère. « Les poids de deux gerbes d'une récolte A, de trois gerbes d'une récolte B, de quatre gerbes d'une récolte C sont supérieurs à une unité de poids. Deux gerbes A valent, en sus de l'unité, une gerbe B. Trois gerbes B valent, en sus de l'unité, une gerbe C, et quatre gerbes C, une gerbe A. Quel est le poids d'une gerbe de chaque récolte ? »

Le système d'équations à résoudre peut s'écrire :

$$2x = 1 + y\,;\,3y = 1 + z\,;\,4z = 1 + x\,;$$

ou :

$$2x - y = 1\,;\,3y - z = 1\,;\,4z - x = 1.$$

Le calculateur chinois dispose sur un échiquier trois colonnes qui vont représenter les trois équations. Sur la première à droite, il place en première ligne deux bâtonnets de couleur ($2x$), en deuxième ligne un bâtonnet noir ($-y$), en quatrième ligne un bâtonnet de couleur : 1 unité.

— 1		2
	3	— 1
4	— 1	
1	1	1

Il procède de façon analogue pour les autres colonnes. Doublant la colonne de gauche et lui ajoutant celle de droite, il arrive à la nouvelle disposition :

		2
— 1	3	— 1
8	— 1	
3	1	1

Triplant la nouvelle colonne de gauche et lui ajoutant la colonne centrale, il obtient la disposition :

		2
	3	— 1
23	—1	
10	1	1

On voit ainsi que 23 $z = 10$, et $z = 10/23$, puis 3 $y - 10/23 = 1$, d'où $y = 11/23$, et 2 $x - 11/23 = 1$, d'où $x = 17/23$.

Cette solution, très remarquable, nécessite que tous les coefficients dans les équations soient des nombres entiers. Elle implique la connaissance des nombres négatifs. L'ouvrage d'où elle est extraite donne d'ailleurs les règles des signes pour les deux opérations fondamentales. Enfin le calculateur tilise les fractions dans leur généralité. En résumé, les mathématiciens chinois travaillaient, pour les systèmes d'équations affines, sur le corps Q des nombres rationnels.

Simple et double fausses positions

On trouve, dans *Neuf Chapitres sur l'art du calcul*, nettement expliquées, les deux règles de la fausse position simple, et de la double fausse position : lorsqu'un problème conduit pour nous à une équation $ax = b$, le calculateur, qui ne dispose pas du calcul littéral, est souvent très gêné pour trouver le coefficient a. S'il connaît le terme b, il effectue, sur une « fausse position » x_0 mise à la place de l'inconnue

x, tous les calculs proposés dans le problème. Il obtient ainsi une valeur b_0 telle que $ax_0 = b_0$. Il ne lui reste plus qu'à résoudre la « proportion » :

$$\frac{x}{x_0} = \frac{b}{b_0} \Leftrightarrow x = x_0 \times \frac{b}{b_0}.$$

Dans d'autres cas plus compliqués, il lui est difficile de calculer les deux coefficients a et b. Une première position x_0 donne $ax_0 - b = r_0$ r_0 est l'erreur. Une seconde position x_1 donne ax_1 $b = r_1$; r_1 est une seconde erreur. Les facteurs a et b ne sont pas connus, mais les quatre nombres x_0, x_1, r_0, r_1 le sont.

Le calcul de x se fait alors par annulation du déterminant :

$$\begin{vmatrix} x & 0 & 1 \\ x_0 & r_0 & 1 \\ x_1 & r_1 & 1 \end{vmatrix}$$

c'est-à-dire que :

$$x = \frac{x_0 r_1 - x_1 r_0}{r_1 - r_0}$$

Bien attestées dans l'ancienne mathématique chinoise, les règles de fausse position sont connues des Arabes et de l'Occident sous le nom d'al-khatayn (la chinoise). Elles existent toujours : c'est l'interpolation linéaire.

Indiquons enfin que les systèmes d'équations affines se rencontrent dans la littérature grecque, principalement chez Diophante d'Alexandrie (IIIe siècle env.). Ce n'est d'ailleurs qu'un aspect mineur de l'œuvre du grand algébriste. Diophante ne procède pas par fausses positions. Il utilise une inconnue, pour laquelle il dispose d'une notation et d'une dénomination.

tes algébristes de la Renaissance

Au XVe siècle, l'Italien Pacioli (1494) et le Français Nicolas Chuquet (1484). entre autres, s'intéressent particulièrement aux systèmes d'équations affines. Le premier de ces deux algébristes utilise parfois une inconnue privilégiée, la *cosa*, et parfois même une seconde, la *quantita.* Cela lui permet la résolution de systèmes à plusieurs inconnues.

Chuquet note l'inconnue 1^1 et résout à notre façon les problèmes affines à une seule inconnue. Pour les problèmes à plusieurs inconnues, en plus des méthodes traditionnelles, il lui arrive d'introduire soit une, soit deux inconnues privilégiées notées alors respectivement 1' et 1'. D'autre part, Chuquet utilise habilement les nombres négatifs. Il rejoint en cela les algébristes chinois et indiens, dépassant de beaucoup les quelques essais timides des savants occidentaux en la matière. Il faut cependant citer, parmi ses précurseurs en ce domaine, Léonard de Pise (XIIIe siècle).

En cette fin du XVe siècle toutefois, la distinction entre systèmes déterminés et indéterminés n'est pas claire. En particulier, on ne voit pas précisément si le problème doit, pour être possible et déterminé, contenir plus, autant ou moins d'équations que d'inconnues.

Le XVIe siècle apporte des progrès appréciables. L'Allemand Michael Stifel (1487-1567) note en 1544 l'inconnue d'un signe particulier analogue à *r,* mais, lorsqu'il se présente d'autres inconnues, il les désigne par les premières lettres de l'alphabet A, B, etc. Les données sont bien entendu numériques (appartenant à Q'). Il est imité, en France, par Jacques Peletier (1517-1582) dans son *Algèbre,* de 1554. Jean Borrel, ou Buteo (1492-1572), est plus net encore en 1559. Soit à résoudre le problème : « Étant donné une somme quelconque, trouver trois nombres dont le premier avec la moitié, le second avec le

tiers, le troisième avec le quart des autres font chacun cette somme. » Il désigne les nombres par A, B, C, et, 17 étant la somme, il écrit :

$$1\,A, \frac{1}{2}B, \frac{1}{2}C\,[17\,;$$
$$1\,B, \frac{1}{3}A, \frac{1}{3}C\,[17\,;$$
$$1\,C, \frac{1}{4}A, \frac{1}{4}B\,[17\,;$$

d'où :

$$2\,A.\,1\,B.\,1\,C\,[34\,;$$
$$1\,A.\,3\,B.\,1\,C\,[51\,;$$
$$1\,A.\,1\,B.\,4\,C\,[68.$$

Procédant alors exactement à la chinoise, il résout très clairement le système.

Lorsque dans un système affine les données sont rationnelles et les inconnues, par leur nature, entières, le système est dit diophantien. II est généralement impossible si le nombre des équations est égal ou supérieur à celui des inconnues. Dans le cas où il y a moins d'équations que d'inconnues, il est indéterminé. La première étude scientifique des systèmes affines diophantiens est due à Bachet de Méziriac, en 1624. Il montre en particulier que, si les entiers a et b sont premiers entre eux, il existe des entiers x et y tels que $ax + by = 1$.

À part cela, le XVIIe siècle apporte peu dans la théorie des équations affines, sinon le développement par Descartes du calcul littéral de Viète. Cependant Leibniz entrevoit le calcul matriciel. Au siècle suivant, le Suisse Gabriel Cramer (1704-I 752) fait la première étude exhaustive des systèmes d'équations affines (1750). Avec le XIXe siècle apparaît le calcul des déterminants, puis le calcul matriciel. Ces problèmes sont à l'origine de l'étude des espaces vectoriels et de toute l'algèbre linéaire.

2. Le second degré

L'histoire des équations quadratiques :

$$ax^2 + bx + c = 0$$

remonte, comme celles des équations affines, à des époques très reculées. La mathématique égyptienne n'a pratiquement rien découvert en ce domaine. Au contraire, l'on doit beaucoup, l'essentiel même, aux Babyloniens.

Premier exemple

Sur une tablette de l'ancien âge babylonien (YBC 4663), on demande de trouver un rectangle, connaissant son demi-périmètre, 6” 30′, et son aire, 7” 30’.

Il s'agit donc de résoudre le systëme :

$$x + y = 6^\circ\,30',$$
$$xy = 7^\circ\,30'.$$

Voici la méthode proposée par le scribe (numération à base 60) : prendre la moitié de la longueur et de la largeur :

$$\frac{x+y}{2}\text{—}\,3^\circ\,15' :$$

élever au carré :

$$\left(\frac{x+y}{2}\right)^2 = 10033'45'' ;$$

en retrancher l'aire :

$$\left(\frac{x+y}{2}\right)^2 - xy = 303'45'' ;$$

prendre la racine carrée : 1” 45’ ; ajouter la demi-somme :

$$3^\circ\,15' + 1^\circ\,45' = 5^\circ \text{ ou } x\,;$$

retrancher :

$$3^\circ\,15' - 1^\circ\,45' = 1^\circ\,30' \text{ ou } y.$$

Second exemple

Problème 7 de la tablette BM 13901. remontant à l'ancien âge babylonien, 1800 environ avant notre ère : « J'ai additionné sept fois le côté de mon carré et onze fois la surface : 6^0 15'. » Soit 11 $x^2 + 7x = 6^0$ 15'.

Solution : « Tu inscriras 7 et 11. Tu porteras 11 à 6^0 15' : 1' $8^0 45'$. Tu fractionneras en deux 7 : 3^0 30'. Tu croiseras 3^0 *30'* et 3^0 30' : $12^0 15'$. À 1' $8^0 45'$ tu ajouteras 1' 21^0, qui est le carré de 9. Tu soustrairas 3^0 30', que tu as croisé, de 9 : tu inscriras 5^0 30'. L'inverse de 11 ne peut être dénoué. Que dois-je poser à 11 qui me donne 5^0 30' : 30', son quotient. Le côté du carré est 30'. »

À savoir : l'équation à résoudre est $ax^2 + bx = c$. On calcule $b^2/4$, puis $b^2/4 + ac$, dont la racine est $\sqrt{(b^2/4 + ac)}$. On forme $\sqrt{(b^2/4 + ac)}$ $b/2$. Le coefficient a n'ayant pas d'inverse dans l'anneau des nombres exprimables en base 60, on divise par a, par tâtonnements, $\sqrt{(b^2/4 + ac)}$ $b/2$. Le quotient est le côté cherché. La numération est sexagésimale : $1' + 60^0$, et $1' = (1/60).\,1^0$.

D'autres exemples. fort nombreux, montrent que le calculateur babylonien sait résoudre toutes les équations quadratiques. Il y a pourtant un obstacle : ce calculateur ne s'exprime pas dans R, corps des réels, mais dans un sous-anneau, celui des nombres exprimables d'une façon finie, en base 60. Pour que $ax^2 + bx + c = 0$ ait des racines, il ne suffit donc pas que $b^2 - 4ac \geqslant 0$, mais il faut encore que cette quantité soit le carré d'un élément de l'anneau et que, de plus, la division finale par a soit possible dans l'anneau.

Les Grecs font, de la résolution des équations du second degré, la base même de toute leur géométrie. Mais, pour pouvoir travailler dans R, ils remplacent les calculs babyloniens par des constructions à la règle et au compas. Pour qu'une équation quadratique ait alors des racines, il suffira que $b^2 - 4ac \geqslant 0$.

Le fait de construire les solutions des problèmes à partir des segments donnés, à la règle et au compas, conduit les géomètres grecs à l'étude des binômes, segments dont les mesures sont de la forme $\sqrt{a} + \sqrt{b}$, a et b rationnels. Cette étude savante n'aboutit guère à des conclusions définitives. Elle occupe cependant une grande partie des *Éléments* d'Euclide et joue un rôle important dans le développement de la théorie des équations. Cependant les algébristes grecs calculent dans $\mathbf{Q}$, plus précisément dans $\mathbf{Q}^+$. Pour eux, une condition supplémentaire s'impose : b^2 4 ac doit être le carré d'un rationnel. Toute l'algèbre diophantienne trouve là son origine. Elle est tenue à manipuler des équations indéterminées où certaines expressions doivent être des carrés parfaits dans Q. L'extraordinaire habileté de Diophante en ce domaine sera un très puissant stimulant pour les mathématiques des XVIe et XVIIe siècles.

Les Arabes et leurs disciples occidentaux jusqu'au XVIe siècle n'apportent rien d'essentiel. La nécessité de calculer dans $\mathbf{Q}^+$ ou dans $\mathbf{R}^+$ les conduit au contraire à distinguer dans les équations quadratiques de multiples cas, assez inutilement. Tout au plus savent-ils que l'équation peut, parfois, admettre deux racines (positives).

3. Équations de degré 3 et 4

Les équations cubiques

Quelques exemples d'équations cubiques apparaissent chez les Babyloniens, mais sans rien de systématique. Archimède

discute *(De lu sphère et du cylindre,* livre second) les problèmes qui, pour nous, conduisent à l'équation cubique générale. Mais sa démarche est purement géométrique et ne peut pas se traduire en algèbre. Le XVe siècle connaît quelques tentatives malheureuses de résolution algébrique. Il était réservé à l'école italienne du XVIe siècle d'apporter la solution définitive. Les trois pionniers sont successivement Scipione del Ferro, Tartaglia et Cardan.

L'équation générale se ramène à des formes telles que $x^3 + px + q = 0$. (Les algébristes n'écrivant que des coefficients numériques et positifs, trois cas sont à distinguer : $x^3 = x + 1$, $x^3 + x = 1$ et $x^3 + 1 = x$.)

La solution trouvée se résume pour nous dans la formule :

$$x = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}.$$

Elle est obtenue en posant $x = u + v$, puis $u^3 + v^3 = \quad q$, $uv = p/3$, d'où $u^3 + v^3 = \quad q$, $u^3v^3 = \quad p^3/27$; u^3 et v^3 sont donc racines d'une équation quadratique.

Cardan comprend aussitôt les difficultés soulevées par cette solution. Lorsque $q^2/4 + p^3/27$ est négatif, les nombres u et v ne peuvent pas être calculés dans R, donc n'existent pas. Or, Archimède a montré que, dans ce cas, l'équation cubique proposée a des racines, et Cardan, en acceptant les racines négatives, sait en outre qu'elle en a trois. Pour lever la difficulté, il introduit timidement, et Bombelli le fera plus nettement en 1572, de nouveaux nombres dits « impossibles » ou « imaginaires ». Ainsi apparaît, pour la première fois, le corps C des nombres complexes.

Le quatrième degré

Un disciple de Cardan, Ferrari, résout l'équation du quatrième degré. Soit par exemple à résoudre $x^4 + 6x^2 + 36 = 60x$. Ajoutons $6x^2$ aux deux membres pour que le premier soit un carré parfait. Il vient $(x^2 + 6)^2 = 6x^2 + 60x$. Formons :

$$(x^2 + 6 + y)^2 = (x^2 + 6)^2 + 2y(x^2 + 6) + y^2.$$

L'équation s'écrit :

$$(x^2 + 6 + y)^2 = (6 + 2y)x^2 + 60x + 12y + y^2.$$

Si le trinôme en x du second membre est un carré parfait $(ax + b)^2$, l'équation se ramènera au second degré : $x^2 + 6 + y = ax + b$. Pour cela, il faut **que :**

$$900 — (6 + 2y)(12y + y*) = 0.$$

Le paramètre y est donc obtenu par la résolution d'une équation cubique. C'est Bombelli qui, en 1572, étend le procédé de Luigi Ferrari à l'équation la plus générale de degré 4.

L'obligation de n'avoir dans les équations que des coefficients positifs rend la démarche de ces auteurs fort pénible.

4. La théorie « générale » des équations

Grâce à l'école italienne, la théorie générale des équations algébriques se précise et ses problèmes principaux se dégagent. Sans suivre chronologiquement son développement historique, on peut s'efforcer d'en mettre en évidence les points importants. L'équation étant mise sous la forme $P(x) = 0$, l'importance du degré du polynôme P, ou degré de l'équation, apparaît d'abord ; en effet, l'équation n'a pas en général une seule racine, comme le voulaient les anciens algébristes, mais peut en avoir jusqu'à n, si n est son degré.

Si a est une racine, alors P(x) est divisible par x a et l'on peut écrire :

$$P(x) = (x - a)Q(x),$$

Q étant un polynôme de degré n 1.

Si l'équation admet exactement n racines, il est possible d'exprimer les coefficients du polynôme P(x) par des fonctions symétriques rationnelles entières des racines.

Exemple du second degré :

$$x^2 - px + q = 0,$$

de racines a et b : alors :

$$(x - a)(x - b) \equiv x^2 - px + q\,;$$
$$p = (a + b)\,;\; q = ab.$$

Exemple du troisième degré :

$$x^3 - px^2 + qx - r = 0,$$

de racines a, b, c ;

$$(x - a)(x - b)(x - c) \equiv x^3 - px^2 + qx - r\,;$$
$$p = a + b + c\,;\; q = ab + ac + bc\,;\; r = abc.$$

Exemple du cinquième degré :

$$x^5 - px^4 + qx^3 - rx^2 + sx - t = 0,$$

de racines a, h, c, d, e ;

$$p = a + b + c + d + e,$$
$$q = ab + ac + ad + ae + bc + bd + be + cd + ce + de,$$
$$r = abc + abd + abe + acd + ace + ade + bcd + bce + bde + cde,$$
$$s = abcd + abce + abde + acde + bcde,$$
$$t = abcde.$$

Ces relations apparaissent déjà chez Viète, dans le seul cas où toutes les racines sont positives, mais c'est Harriot, en 1630, dans ses œuvres posthumes, et surtout Albert Girard, en 1629, qui leur donnent toute leur extension. Girard, d'autre part et il sera suivi par Newton exprime les sommes des puissances des racines en fonction des coefficients :

$$a + b + c + d + e = p,$$
$$a^2 + b^2 + c^2 + d^2 + e^2 = p^2 - 2q,$$
$$a^3 + b^3 + c^3 + d^3 + e^3 = p^3 - 3pq + 3r,$$
$$a^4 + b^4 + c^4 + d^4 + e^4 = p^4 - 4p^2q + 4pr + 2q^2 - 4s.$$

L'étude des fonctions symétriques des racines se développe considérablement au XVIII[e] siècle avec Waring, au XIX[e] avec Cauchy, etc.

Ces belles relations ne sont évidemment établies, chez Viète, que lorsque toutes les racines sont positives, et, pour tout algébriste, que si elles existent. Tout dépend du sens donné au mot «existence ». Pour Jean de Beaugrand par exemple, vers 1638, exister est synonyme d'« appartenir à l'ensemble R des réels ». Pour Girard, on peut admettre des « solutions impossibles » pour la « certitude de la règle générale et pour son utilité ». Pour Descartes, en 1637, « les racines ne sont pas toujours réelles, mais quelquefois seulement imaginaires, c'est-à-dire qu'on peut bien toujours en imaginer autant que j'ai dit en chaque équation, mais qu'il n'y a quelquefois aucune quantité qui corresponde à celle qu'on imagine ». Il semble bien que ce soit Peter Roth de Nuremberg qui ait, le premier, en 1608, énoncé cet aphorisme hardi : « Une équation a autant de racines qu'il y a d'unités dans son degré. » Cette conclusion est une conséquence des principes énoncés par Bombelli, sans leur être identique. Cet algébriste italien introduit $\sqrt{-1}$, qu'il adjoint aux nombres réels, créant ainsi le corps C des nombres complexes. Il peut alors retrouver les racines réelles de l'équation cubique dans le cas dit « irréductible ». Roth, suivi par Girard et Descartes, puis par la grande majorité des mathémati-

ciens, décide, très arbitrairement, l'existence d'êtres fictifs, n'appartenant pas à l'ensemble R des nombres réels (et dont on ignore s'ils appartiennent ou non à C), et étend à ces êtres les algorithmes classiques de calcul. Jusqu'en 1746, on maniera ainsi des êtres imaginaires, sans trop savoir quelle pourrait bien être leur structure. Cependant la conviction se répandait de plus en plus qu'ils étaient de la forme $a + b\sqrt{-1}$. C'est ce que d'Alembert établit cette année-là, en s'appuyant sur le calcul infinitésimal et la géométrie analytique et en admettant le principe d'existence des n racines d'une équation de degré n. Daviet de Foncenex, Lagrange, Laplace améliorèrent cette démonstration, mais en se fondant toujours sur le même principe. Gauss, en 1799, qualifia de cercle vicieux cette démarche et il fournit enfin plusieurs preuves rigoureuses du « théorème fondamental de l'algèbre », ou « théorème de d'Alembert ».

On sait aujourd'hui que l'attitude des géomètres du XVIIIe siècle peut se justifier en ce domaine, et cela, grâce à la théorie des congruences de Gauss et à l'introduction par Galois du corps de rupture. Précisons en quelques mots. Soit un corps K, commutatif, et un polynôme P(x), indécomposable sur K en un produit de deux polynômes (premiers sur K). Alors l'anneau des polynômes construits sur K se subdivise en classes d'équivalences : deux polynômes R(x) et S(x) sont équivalents si leur différence est divisible par P(x). On démontre que ces classes d'équivalence forment un corps commutatif. Désignant par a la variable désignée par x jusqu'ici, tous les éléments du corps sont des polynômes en a à coefficients dans K.

Dans ce nouveau corps, $P(a) \equiv 0$; autrement dit, P(x) admet la racine a et est divisible par $x - a$.

Exemple. En prenant, pour P(x), $x^2 + 1$, et, pour K, le corps R des réels, et en notant i ce que nous notions a, on trouve le corps des complexes où $x^2 + 1 = (x + i)(x - i)$. Cette construction du corps des nombres complexes a été proposée par Cauchy.

De ce point de vue, la démarche des précurseurs de Gauss était correcte. Le seul point à établir était de montrer que, pour K = R, P(x) premier quelconque, le corps de rupture est un sous-corps de C.

5. Résolution numérique

Pour une équation P(x) = 0, les questions suivantes se sont posées naturellement : Combien a-t-elle de racines réelles ? Combien d'imaginaires ? Combien de positives, de négatives ? Quelles sont les valeurs approchées, au 1/10, au 1/100, au 1/1000 près, etc. de ces diverses racines?

Descartes, dans sa *Géométrie* de 1637, déclare, sans preuves, que l'équation P(x) = 0 peut avoir autant de racines réelles positives qu'il y a de changements de signe dans les monômes du premier membre ordonné, et autant de négatives qu'il y a de permanences.

Ainsi $x^4 + x^3 - x^2 + x + 1 = 0$ peut avoir deux racines positives et deux racines négatives. En fait, cette équation n'a aucune racine positive et deux racines négatives, les deux autres étant imaginaires conjuguées. On énonce aujourd'hui le théorème de Descartes comme il suit : Dans une équation quelconque, à coefficients réels, le nombre des racines positives ne surpasse pas le nombre des variations de signe du premier membre ; et, quand il est moindre, la différence est toujours un nombre pair.

En 1690, Rolle (16.52-1719) énonce dans son *Algèbre* une proposition que l'on peut exprimer ainsi : Soit P(x) = 0, formons l'équation P'(x) = 0, P' étant le polynôme dérivé du polynôme P. Entre deux racines de la première équation, il existe au moins une racine de la seconde.

Le théorème de Budan (18 11) se rattache au même ordre d'idées : « Étant donné une équation P(x) = 0 de degré m, si dans les $(m + 1)$ fonctions P(x), P' (x), P"(x), . ..) $P^{(m)}(x)$ où chacune est la dérivée de la précédente, on substitue à x deux nombres a et β $(a < \beta)$, et, si après chaque substitution on compte les variations de signe que présente la suite des résultats, le nombre des racines de P(x) = 0 comprises entre a et β ne surpasse jamais celui des variations perdues de a à β, et, quand il est moindre, la différence est toujours un nombre pair. »

Le théorème de Sturm (1829) est le résultat le plus précis qui ait été obtenu dans ce domaine. Soit P(x) = 0 l'équation proposée. On divise P par le polynôme dérivé P'. Soit P_2 le reste euclidien changé de signe. Divisons P' par P_2, et soit P_3 le reste changé de signe, etc. Si m est le degré de P, *supposé sans racines multiples,* considérons la suite P, P', $P_2, P_3, \ldots P_m$. Soit alors, comme dans le théorème de Budan, a et β $(a < \beta)$ deux nombres donnés. Formons P(a), P'(a), ... P,,,(a), et de même $P(\beta), P'(\beta), \ldots P_m(\beta)$.
Le nombre des racines de l'équation comprises entre a et β est précisément égal à l'excès du nombre ûes variations de signe que présente la première suite sur celui que présente la seconde.

Les théorèmes précédents et quelques autres analogues permettent la séparation des racines de l'équation. C'est-à-dire que, pour chaque racine réelle, on arrive à trouver deux nombres a et β entre lesquels il n'existe que cette racine de l'équation. À partir de là, on peut appliquer les méthodes générales de résolution d'une équation f(x) = 0 pour obtenir des valeurs approchées des racines.

6. La résolution algébrique des équations

Par cette expression, on entend traditionnellement la résolution des équations au moyen de radicaux carrés, cubiques, etc.

On a vu que sont résolubles par ce procédé les équations de degrés 2, 3 et 4. Après les succès de l'école italienne au XVIe siècle, les mathématiciens se sont attachés à trouver des formules de résolution analogues pour les degrés suivants, singulièrement pour le cinquième. Parmi les recherches les plus remarquables en ce domaine, on peut citer celles de Tschirnhaus (16.5 l-1 708). 11 s'efforce, en 1689, par un changement de variable, de ramener toute équation algébrique à une équation binôme. Plus précisément, soit P(x) = 0 une équation de degré n. Posons y = Q(x), Q étant un polynôme de degré $n - 1$ à coefficients indéterminés. On élimine x entre les deux équations P(x) = 0 et Q(x) y = 0, et l'on détermine les coefficients du second polynôme de façon à faire disparaître, dans l'équation résultante en y, certains ou tous les termes intermédiaires. Si la méthode de Tschirnhaus réussissait toujours, toute équation serait aigébriquement résoluble. A-u XVIIIe siècle, Euler et Bezout ont étudié le même problème par des procédés analogues.

Un mémoire de Vandermonde, lu en novembre 1770, devait inaugurer une ère nouvelle. Kronecker n'a pas craint d'affirmer que l'essor moderne de l'algèbre commençait avec ce mémoire. Vander-

monde y apparaît comme le précurseur et le premier ouvrier de la théorie des substitutions, distinguant, avant Gauss et Abel, les fonctions cycliques invariantes par une permutation circulaire déterminée et décomposant les fonctions symétriques en fonctions cycliques, Naturellement, il n'aboutit pas pour les degrés 5 et 6, mais il montre combien il serait prématuré de conclure à l'impossibilité de la résolution des équations générales de degré supérieur à 4. Puis il note que, si sa méthode échoue pour ces équations générales, elle réussit pour des équations particulières dont les racines sont liées par certaines relations et il prend pour exemple $x^{11} - 1 = 0$, dont il exprime les solutions au moyen de racines carrées et de racines cinquièmes.

Avec Vandermonde apparaît ainsi la notion de substitution dans un ensemble fini, celui des racines d'une équation algébrique, notion que devaient approfondir les algébristes ultérieurs, et dont l'étude aboutira, avec Galois, au concept nouveau de groupe fini.

Les idées développées par Vandermonde se trouvent encore, indépendamment d'ailleurs, dans l'important mémoire de Joseph Lagrange, lu en 1771 : *Réflexions sur la résolution algébrique des équations.*

Gauss, dans ses *Disquisitiones arithmeticae* (1801), explicite les remarques de Vandermonde sur les équations binômes $x^n - 1 = 0$, les appuie solidement sur les propriétés arithmétiques de l'exposant et montre notamment, à partir de l'équation $x^{17} - 1 = 0$, la possibilité d'inscrire dans le cercle, à la règle et au compas, un polygone régulier de 17 côtés.

Les travaux de Vandermonde, Lagrange et Gauss attirèrent en particulier l'attention des géomètres sur les fonctions entières de plusieurs variables et sur les changements qu'elles éprouvent dans une permutation de ces variables. Lagrange démontra que le nombre des valeurs d'une fonction de n lettres est toujours un diviseur de $n\,!$, produit des n premiers entiers. Ruffini (1765-1822) établit en 1799 que si une fonction de cinq variables a moins de cinq valeurs distinctes, elle ne peut en avoir plus de deux. Si ce théorème n'établit pas l'impossibilité de la résolution algébrique de l'équation générale du cinquième degré, il prouve du moins l'impossibilité de former une équation auxiliaire ou résolvante de degré inférieur à 5. Cauchy généralise : Si une fonction de n lettres a moins de p valeurs distinctes (p plus grand nombre premier contenu dans n), elle ne peut en avoir plus de deux (1815).

Ces diverses propositions nécessitaient déjà l'étude de la structure de l'ensemble des substitutions de n lettres, en particulier celle des substitutions circulaires.

Abel (1802-1829), dans le même ordre d'idées, avait établi qu'une fonction de 5 lettres ayant cinq valeurs distinctes est symétrique par rapport à 4 lettres. Appliqué à la résolution algébrique des équations, ce théorème montrait que, si l'on cherche à faire dépendre la résolution de l'équation générale du cinquième degré de celle d'une résolvante de ce même degré, on revient à la transformation de Tschirnhaus. Ruffini énonça en 1813 l'impossibilité de la résolution algébrique de cette équation générale du cinquième degré. En 1824 et 1826, Abel apporta sur ce point des arguments plus probants. De plus, généralisant l'analyse de Vandermonde et de Gauss pour les équations binômes, il établit que, si deux racines d'une équation irréductible sur le corps des coefficients sont telles que l'une puisse s'exprimer rationnellement par l'autre, l'équation est

résoluble par radicaux si son degré est un nombre premier, et que, s'il n'est pas premier, sa résolution dépend d'équations de degrés moindres que le sien.

Abel, dans son étude de l'équation du cinquième degré, se servait du fait que les quantités successives dont il faudrait, dans cette résolution, extraire les racines n-ièmes doivent s'exprimer rationnellement en fonction des racines cherchées. Ce point présente des difficultés. Galois (1811-1832) procède par une démarche différente (1830). En appelant *groupe* d'une équation, sur un corps donné, le groupe des permutations de ses racines qui laissent inchangées les expressions polynomiales des racines dont la valeur appartient à ce corps, il montre que, dans une résolution par radicaux, et dans les réductions successives que subit, au cours des calculs, le groupe de l'équation, chaque nouveau groupe est un sous-groupe invariant du précédent. Or, le groupe des substitutions de cinq lettres n'a pas de sous-groupe invariant. Donc la résolution algébrique de l'équation générale du cinquième degré est impossible. De plus, sa méthode lui permit (1831) de montrer que, pour qu'une équation irréductible de degré premier soit soluble par radicaux, il faut et il suffit que, deux quelconques des racines étant données, les autres s'en déduisent rationnellement.

À un point de vue plus élémentaire, mais historiquement très important, signalons le mémoire du mathématicien français Pierre Laurent Wantzel (1814-1848) : « *Recherches sur les moyens de reconnaître si un problème de géométrie peut se résoudre par la règle et le compas* » (1837). Wantzel y montre pour la première fois, d'une façon irréfutable, que, si un problème de géométrie conduit à une équation de troisième degré, indécomposable sur le corps de ses coefficients, ce problème n'est pas résoluble à la règle et au compas. Ainsi se trouve justifiée la distinction grecque entre les « problèmes solides », comme ceux de la duplication du cube et de la trisection de l'angle, et les « problèmes plans », traitables à la règle et au compas.

JEAN ITARD

Bibliographie

N. BOURBAKI, *Éléments d'histoire des mathématiques*, Masson, 1984 / E. M. BRUINS & M. RUTTEN, *Textes mathématiques de Suse*, Paris, 1961 / E. DEHN, *Algebric Equations. An Introduction to the Theories of Lagrange and Galois*, repr. Dover Publ., New York, 1960 / P. DREDON & J. ITARD, *Mathématiques et mathématiciens*, Magnard, nouv. éd., 1980 / J.-C. MARTZLOFF, *Histoire des mathématiques chinoises*, Masson, 1987 / C. MUTAFIAN, *Équations algébriques et théorie de Galois*, Vuibert, Paris, 1980 / J.-A. SERRET, *Cours d'algèbre supérieure*, 2 vol., repr. 1866, J. Gabay, 4ᵉ éd. 1992 / J. TROPFKE, *Geschichte der Elementarmathematik*, 4ᵉ éd., vol. 1 : *Arithmetik und Algebra*, de Gruyter, Berlin-New York, 1980 / B. L. VAN DER WAERDEN, *A History of Algebra*, Springer-Verlag, New York, 1990 / *Science Awakening*, vol. 1 : *Egyptian, Babylonian und Greek Mathematics*, repr. 1954, Scholar's Bookshelf, Princeton Junction (N. J.), 1988 / F. VIETE, A. GIRARD & F. DE BEAUNE, *The Early Theory of Equations : on their Nature and Constitution*, Golden Hind Press, Fairfield (Conn.), 1986.

ÉQUATIONS AUX DÉRIVÉES PARTIELLES → DÉRIVÉES PARTIELLES ÉQUATIONS AUX

ÉQUATIONS DIFFÉRENTIELLES → DIFFÉRENTIELLES ÉQUATIONS

ÉQUATIONS DIOPHANTIENNES → DIOPHANTIENNES ÉQUATIONS

ÉQUATIONS INTÉGRALES → INTÉGRALES ÉQUATIONS

ERGODIQUE THÉORIE

Ergodique vient du mot grec ἔργον qui signifie travail. C'est en effet d'un problème de mécanique que la théorie ergodique est issue. À l'origine se trouve une hypothèse de la théorie cinétique des gaz, audacieusement posée par L. Boltzmann en 1885, qui permettait aux physiciens de résoudre une difficulté liée à l'étude des systèmes mécaniques à un très grand nombre de particules. L'importance de cette hypothèse, confirmée expérimentalement dans de nombreux cas, conduisit les mécaniciens à en chercher une justification théorique et ce sont les diverses tentatives faites dans cette voie qui marquent les débuts de la théorie ergodique.

Après les résultats fondamentaux, obtenus par J. von Neumann et G. D. Birkhoff en 1931 à quelques semaines d'intervalle, la théorie ergodique s'est développée au sein de la mathématique dans des directions diverses : analyse fonctionnelle et théorie des groupes ; calcul des probabilités et plus précisément processus markoviens ; théorie de l'information, etc. Les méthodes ergodiques ont permis d'exposer différemment certains problèmes et de donner des prolongements nouveaux à ces branches de la mathématique. L'étude de la théorie ergodique suppose la connaissance de la théorie de la mesure (cf. INTÉGRATION ET MESURE).



1. Le modèle de Poincaré et l'hypothèse ergodique

Pour expliquer l'hypothèse ergodique, il est commode d'avoir recours à un modèle très simple imaginé par H. Poincaré. Supposons un liquide en mouvement stationnaire dans un récipient Ω de forme invariable et complètement rempli. Si une molécule du liquide occupe la position ω_0 à l'instant 0 et ω_t à l'instant t, on peut décrire le passage de l'instant 0 àl'instant t et, plus généralement, de l'instant s à l'instant $s + t$, au moyen d'une transformation ponctuelle θ_t opérant dans Ω, pour laquelle :

$$\omega_t = \theta_t(\omega_0)$$

et :

$$\omega_{s+t} = \theta_t(\omega_s),$$

et cela pour toutes les molécules de liquide. Il va de soi que ces transformations forment un groupe, c'est-à-dire que :

$$\theta_{s+t} = \theta_s \circ \theta_t, \quad (1)$$

quels que soient $s, t \in \mathrm{R}$; ou bien, si l'on se désintéresse du passé, un semi-groupe en ne considérant la condition (1) que pour des valeurs positives ou nulles s et t. Pour $t = 0$, θ_0 désigne la transformation identique de Ω. On peut envisager une simplification supplémentaire en se limitant aux instants discrets . . . . 2, 1, 0, 1, 2, et,

en posant $\theta_1 = \theta$, se ramener à l'étude du groupe G à un générateur θ :

$$\mathbf{G} = \{\theta^n \mid n \in \mathbf{Z}\}$$

ou bien du semi-groupe :

$$\{\theta^n \mid n \in \mathbf{N}\}.$$

En outre, l'incompressibilité du liquide amène à poser la condition suivante : Si E est un volume partiel de Ω et si :

$$\theta^{-1}\mathbf{E} = \{\omega \mid \theta\omega \in \mathbf{E}\},$$

alors les mesures de E et de θ^{-1}E sont égales ; autrement dit, la transformation θ conserve la mesure. Pour un point donné $\omega \in \Omega$, l'ensemble :

$$\{\theta^n\omega \mid n \in \mathbf{Z}\ (\text{ou}\ \mathbf{N})\}$$

est appelé *trajectoire* de ω. On dit qu'un point ω est *topologiquement infiniment récurrent* si tout voisinage de ce point possède une infinité de points de sa trajectoire. On peut alors énoncer le théorème de récurrence qui peut être considéré comme le premier théorème ergodique et qui fut établi par H. Poincaré en 1890.

Théorème de Poincaré. Presque tout point de Ω est topologiquement infiniment récurrent.

En vérité, cela n'est pas exactement l'énoncé donné par le célèbre géomètre qui ne pouvait pas faire usage à cette époque de la théorie de la mesure de Lebesgue, théorie qui permit quelques années plus tard de prouver le théorème de Poincaré.

Revenant au problème général, considérons un système mécanique S constitué par N particules. L'état de S à chaque instant est déterminé par la connaissance des 3 N coordonnées des particules et des 3 N composantes de leurs vitesses. Ces états peuvent ainsi être représentés par des points ω de l'espace $\mathbf{R}^{6N}$ appelé espace de phase. Il peut être aussi plus commode de substituer à ces 6 N coordonnées 6 N autres paramètres qu'on ne précisera pas ici. Le système S évolue sous l'action de forces extérieures et des actions mutuelles des particules, et cette évolution est régie par un système d'équations différentielles qui permet de déterminer à partir d'un état initial ω_0 l'état ω_t du système à l'instant t. Si S est conservatif, les trajectoires des points ω sont portées par des variétés plongées dans $\mathbf{R}^{6N}$ et chacune de celles-ci correspond à une valeur constante de l'énergie de S, ce qui explique l'emploi du terme ergodique. L'une quelconque de ces variétés sera pour nous le récipient Ω du modèle de Poincaré. L'invariance de la mesure, lorsque les paramètres décrivant l'état de S sont convenablement choisis, est assurée par un résultat général dû à Liouville. Le physicien qui observe l'évolution du système s'intéresse à des mesures portant sur les valeurs de fonctions de l'état de S (observables). Soit f l'une d'elles, $f : \Omega \rightarrow$ R. Si l'on tient compte d'une part du grand nombre de particules et du fait que, dans un intervalle de temps très petit pris pour unité, se produit un nombre très élevé de microphénomènes (collision de particules, etc.), si l'on tient compte d'autre part de la durée qu'exige une mesure, on s'aperçoit que le physicien ne peut pas comparer ses mesures aux valeurs théoriques instantanées f(o) mais plutôt aux moyennes temporelles :

$$(2) \qquad \frac{1}{n}\sum_{k=0}^{n-1} f(\theta^k\omega)$$

portant sur n instants consécutifs 0, 1, 2, ... n 1 et cela pour de très grandes valeurs de n. Ensuite, en supposant même que l'état initia1 ω soit parfaitement connu, la détermination des états successifs $\theta\omega, \theta^2\omega$,

..., $\theta^{n-1}\omega$ exigerait l'intégration du système différentiel mentionné plus haut, calcul pratiquement impossible à effectuer. Il faut donc imaginer un autre moyen d'atteindre les quantités théoriques (2) et c'est là qu'intervient l'hypothèse ergodique. Cette hypothèse postule l'égalité des moyennes de phase :

$$\frac{1}{m(\Omega)}\int_\Omega f dm$$

et des moyennes (2) pour n assez grand, Ω étant la variété associée aux données de l'expérience.

2. Les théorèmes de G. D. Birkhoff et de J. von Neumann

Qn va mamtenant formaliser le problème ergodique. On se donne un espace compact Ω et une mesure de Radon positive m sur Ω (cf. INTÉGRATION ET MESURE ; on peut se placer dans des situations plus générales, mais on n'a pas jugé utile de le faire ici), qui est aussi une probabilité $m(\Omega) = 1$. On se donne aussi une transformation mesurable $\theta : \Omega \rightarrow \Omega$ et on suppose que θ conserve la mesure, c'est-à-dire que θ vérifie la condition :

(3) $$m(\theta^{-1}E) = m(E)$$

pour tout ensemble mesurable E. Cette condition entraîne que $\theta^{-1}E$ est négligeable si E est négligeable, c'est-à-dire m(E) = 0.

Théorème de Birkhoff. Soit f une fonction complexe et intégrable sur Ω ; la suite des moyennes de Cesaro :

$$\frac{1}{n}\sum_{k=0}^{n-1} f(\theta^k\omega)$$

converge presque partout sur Ω vers une fonction intégrable $\tilde{f}$; cette fonction $\tilde{f}$ est θ-invariante (c'est-à-dire $\tilde{f} = \tilde{f}\circ\theta$) et enfin :

$$\int_A f dm = \int_A \tilde{f} dm,$$

quel que soit l'ensemble mesurable A invariant, c'est-à-dire tel que $A = \theta^{-1}A$.

Dans le cas particulier où la condition (E) suivante est satisfaite : les seuls ensembles invariants sont modulo les ensembles négligeables, l'ensemble Ω et l'ensemble vide, les fonctions invariantes sont les fonctions constantes presque partout (p.p.), et le théorème de Birkhoff donne l'égalité :

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} f\circ\theta^k = \int_\Omega f dm \ (p.p.) ;$$

autrement dit le système (Ω, m, θ) vérifie l'hypothèse ergodique si la condition (E) est remplie. On dit dans ce cas que θ est *transitivement métrique* ou encore que θ est *ergodique.*

Quelques semaines avant que G. D. Birkhoff eût donné son résultat, J. von Neumann avait établi le théorème suivant en faisant les mêmes hypothèses que pour le théorème de Birkhoff.

Théorème de von Neumann. Soit f une fonction complexe sur Ω, de carré intégrable ; la suite des fonctions :

$$\frac{1}{n}\sum_{k=0}^{n-1} f\circ\theta^k$$

converge en moyenne quadratique vers une fonction $\tilde{f}$ de carré intégrable et θ-invariante ; autrement dit :

$$\lim_{n\to\infty}\int_\Omega \left|\frac{1}{n}\sum_{k=0}^{n-1} f\circ\theta_k - \tilde{f}\right|^2 dm = 0$$

et :

$$\tilde{f} = \tilde{f}\circ\theta \quad (p.\,p.).$$

Il n'est pas question de donner ici les démonstrations de ces théorèmes, mais il est utile d'ajouter quelques indications sur ces preuves pour montrer en particulier les liens existant entre la théorie ergodique et l'analyse fonctionnelle.

Dans ces théorèmes intervient une transformation agissant non pas sur les points de Ω mais plutôt sur les fonctions définies sur Ω. Il s'agit de l'application :

$$f \mapsto Tf = f \circ \theta.$$

Pour éviter des difficultés techniques, on ne distinguera pas des fonctions égales presque partout ; et, d'ailleurs, l'hypothèse d'invariance de la mesure entraîne que, si f est négligeable, c'est-a-dire

$$\int |f| \, dm = 0,$$

$f \circ \theta$ l'est aussi.

Cela posé, soit L^p l'espace des fonctions complexes sur Ω de puissances pièmes intégrables :

$$\int |f|^p \, dm < + \infty,$$

p étant un nombre réel donné $1 \leqslant p < + \infty$. On sait que L^2 est muni d'une structure d'espace de Hilbert où le produit hermitien de deux éléments f_1 et f_2 est défini par :

$$< f_1, f_2 > = \int_\Omega f_1 . \overline{f}_2 \, dm ;$$

$\overline{f}_2$ est la conjuguée complexe **de** f_2 et la norme d'un élément f est :

$$\|f\|_2 = (< ff >)^{1/2}.$$

Alors il est facile de voir que T est un opérateur linéaire et unitaire sur L^2. Cela étant, et sans entrer dans les détails, la démonstration du théorème de von Neumann repose sur le fait qu'il existe dans L^2 deux sous-espaces fermés et orthogonaux $\mathcal{J}$ et $\mathcal{H}$ tels que tout $f \in L^2$ s'écrive d'une seule manière : $f = g + h$, avec $g \in \mathcal{J}$ et $h \in \mathcal{H}$, où $\mathcal{J}$ est le sous-espace des invariants ($g \in \mathcal{J} \Leftrightarrow g = Tg$) et $\mathcal{H}$ est l'adhérence du sous-espace image de I -T.

On ne dira rien de la preuve donnée par Birkhoff de son théorème, mais on donnera une indication sur la démonstration proposée par Yosida et Kakutani. Elle est fondée sur le lemme suivant, nommé par ces auteurs « théorème ergodique maximal ».

Soit $f \in L^1$ et E l'ensemble des ω pour lesquels l'une au moins des sommes :

$$\sum_{k=0}^{n} f(\theta^k \omega)$$

est positive ; alors :

$$\int_E f dm \geqslant 0.$$

Illustrons ce qui précède par deux exemples.

a) Ω est le tore à une dimension, $\Omega = \mathbf{R}/\mathbf{Z}$. L'application θ est définie par $\theta\omega = \omega + a$ où a est la classe d'équivalence d'un nombre irrationnel. La mesure m sur Ω est induite par la mesure de Lebesgue sur R. On peut voir facilement ici que θ conserve la mesure et que θ est ergodique. On a ainsi :

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(\omega + ka) = \int_\Omega f \, dm \quad (p. p.),$$

quel que soit $f \in L^1$. Ce résultat avait été démontré directement par Khintchine sans utiliser le théorème de Birkhoff.

b) Ω est l'intervalle [0, 1] muni de la mesure de Lebesgue. Chaque réel $\omega \in [0, 1]$ s'écrit dans le système décimal, au moins d'une manière, $\omega = 0, a_1 a_2 a_3$

où les a_i sont des entiers compris entre 0 et 9. Posons alors :

$$\theta\omega = 0, a_2 a_3 a_4 \ldots,$$

en observant que, dans tous les cas, $\theta\omega$ est bien défini par la donnée de ω. Il est moins facile ici de vérifier que θ conserve la mesure et que θ est ergodique. Admettons-le et prenonsfégale à la fonction caractéristique de l'intervalle $[q/10, q/10 + 1/10[$, avec $q \in \{0, 1, 2, \ldots . 8\}$.

On a $f(\omega) = 1$. si $a_1 = q$ *ou* si $a_1 = q -$ 1 et a, = a_3 = = 9. Le théorème de Birkhoff affirme, dans ces conditions, que :

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(\theta^k \omega) = \frac{1}{10} \ (p. \text{ p.}).$$

Ce résultat, découvert par É. Borel, exprime que, pour presque tout réel x, chaque chiffre admet dans la suite des décimales du nombre x la même fréquence limite $1/10$.

3. Propriétés de mélange

Revenons au modèle de Poincaré et supposons que le liquide enfermé dans le récipient Ω soit, suivant une image de Halmos, un mélange de vermouth et de gin dans les proportions de $9/10$ de gin et $1/10$ de vermouth. Le récipient Ω est un shaker que l'on agite pour confectionner un cocktail. Chaque mouvement d'agitation du shaker s'effectue aux instants $1, 2, \ldots . n$, ... Si B est la partie de Ω occupée initialement par le vermouth, alors, pour toute autre partie mesurable A du shaker, le rapport entre la quantité de vermouth contenue dans A et la quantité totale de vermouth est, à l'instant n,

$$\frac{m(\theta^{-n} \mathrm{A} \cap \mathrm{B})}{m(\mathrm{B})}$$

ou encore, en désignant par 1_A la fonction caractéristique de l'ensemble A,

$$\frac{1}{m(\mathrm{B})} \int_\mathrm{B} 1_\mathrm{A}(\theta^n \omega)\, dm(\omega).$$

La moyenne arithmétique de ces rapports pris aux instants 0, 1, ... $n - 1$ est donc :

$$\text{(4)} \qquad \frac{1}{m(\mathrm{B})} \int_\mathrm{B} \left(\frac{1}{n} \sum_{k=0}^{n-1} 1_\mathrm{A}(\theta^k \omega) \right) dm(\omega).$$

On suppose toujours $m(\Omega) = 1$. Si θ est ergodique, la suite (4) converge vers m(A). Autrement dit, en moyenne, la suite :

$$\frac{m(\theta^{-n} \mathrm{A} \cap \mathrm{B})}{m(\mathrm{B})}$$

converge vers m(A). Cela est, bien cntcndu, rćalisć a fortiori si :

$$\lim_{n \to \infty} \frac{m(\theta^{-n} \mathrm{A} \cap \mathrm{B})}{m(\mathrm{B})} = m(\mathrm{A}),$$

condition qui exprime qu'après un temps assez long toute partie telle que A contient la même proportion de gin et de vermouth qui se trouvent ainsi parfaitement mélangés.

On pose alors les définitions suivantes :

– La transformation θ est dite *fortement mélangeante* si, pour tout couple de parties mesurables A et B de Ω, on a :

$$\lim_{n \to \infty} m(\theta^{-n} \mathrm{A} \cap \mathrm{B}) = m(\mathrm{A}) m(\mathrm{B}).$$

La transformation θ est dite *faiblement mélangeante* si, pour tout couple de parties mesurables A et B de Ω, on a :

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} |m(\theta^{-k} \mathrm{A} \cap \mathrm{B}) - m(\mathrm{A}) m(\mathrm{B})| = 0.$$

Il est clair que toute transformation fortement mélangeante est, a fortiori, faiblement mélangeante et que toute transformation faiblement mélangeante est, a fortiori, ergodique.

La condition de mélange faible est particulièrement intéressante du point de vue analytique. Reprenons l'opérateur T défini au chapitre 2. On peut définir le spectre de T (cf. théorie SPECTRALE). Disons que T a un spectre continu si 1 est la seule valeur propre de T et de plus valeur propre simple. On peut alors prouver que θ est faiblement mélangeante si, et seulement si, T a un spectre continu.

4. Systèmes dynamiques

On ne donnera pas de définition générale et on se limitera aux systèmes (Ω, m, θ) ayant les propriétés énoncées au début du paragraphe 2. On appelle un tel triplet $\mathcal{S} = (\Omega, m, \theta)$ un système dynamique. Soit $\mathcal{S} = (\Omega', m', \theta')$ un autre système dynamique. On dira que S' est image homomorphe de $\mathcal{S}$ s'il existe une injection mesurable $\varphi : \Omega \rightarrow \Omega'$ telle que $\varphi \circ \theta = \theta' \circ \varphi$ et m' $= \varphi(m)$. Si φ est bijective et si chacun des systèmes $\mathcal{S}$ et S' est image homomorphe de l'autre par φ et φ^{-1}, $\mathcal{S}$ et S' sont dits *spatialement isomorphes*. Cela étant, on peut poser la question suivante : Deux systèmes dynamiques donnés sont-ils isomorphes ? Pour y répondre, il est bon de rechercher les invariants d'un système dynamique $\mathcal{S}$, c'est-i-dire les objets attachés à $\mathcal{S}$ qui ne varient pas dans un isomorphisme spatial. Dans le chapitre 2, on a associé à la transformation θ un opérateur unitaire T dans $L^2(m)$. Il est alors facile de vérifier que les valeurs propres de T sont des invariants de S.

Un autre invariant fondamental des systèmes dynamiques est *l'entropie ou* invariant de Kolmogoroff-Sinaï qui peut se définir de la façon suivante : Désignons par χ la fonction réelle continue et positive sur $[0, 1]$, telle que $\chi(x) = -x \log x$, pour $0 < x \leqslant 1$; à toute partition mesurable finie :

$$\Pi = (\Omega_i)_{1 \leqslant i \leqslant n}$$

de Ω, faisons correspondre le nombre $H(\Pi)$ défini par :

$$H(\Pi) = \sum_{i=1}^{n} \chi(m(\Omega_i)).$$

On observe que, pour tout entier k,

$$\theta^{-k}\Pi = (\theta^{-k}\Omega_i)_{1 \leqslant i \leqslant n}$$

est une autre partition de Ω et l'invariance de la mesure entraîne que :

$$(5) \qquad H(\theta^{-k}\Pi) = H(\Pi).$$

La fonction χ est concave et l'on en déduit, pour deux partitions P et P', que :

$$(6) \qquad H(\Pi \vee \Pi') \leqslant H(\Pi) + H(\Pi'),$$

en notant 11 V II' la partition engendrée par II et II'. Les propriétés (5) et (6) permettent de prouver l'existence de la limite :

$$\overline{H}(\Pi) = \lim_{k \to \infty} \frac{1}{k} H(\Pi \vee \theta^{-1}\Pi \vee \ \vee \theta^{-k+1}\Pi),$$

et aussi l'inégalité :

$$\overline{H}(\Pi) \leqslant H(\Pi).$$

On pose alors :

$$\hat{H} = \sup_{\Pi} \overline{H}(\Pi),$$

la borne supérieure étant prise sur l'ensemble des partitions mesurables finies Π dc Ω. Le nombre $\hat{H}$ (éventuellement $+ \infty$) est l'entropie du système S. Ajoutons que cette notion est très voisine de celle qui est utilisée par Boltzmann dans la théorie cinétique des gaz et qu'elle a été l'objet de profonds et difficiles travaux de Sinaï qui, par là. a fait un pas important vers la

solution du problème fondamental de la théorie ergodique (cf. chap. 1).

Pour terminer, donnons un exemple. Soit A = {a,, a_2, ... a_n} un ensemble à n éléments ; munissons-le de la topologie discrète et de la mesure de probabilité équirépartie :

$$p(\{a_i\}) = 1/n,$$

pour i = 1, 2, ..., n. Posons $\Omega = A^Z$; chaque élément de Ω est une suite infinie dans les deux sens $\omega = (..., \omega_{-1}, \omega_0, \omega_1, ...)$ d'éléments de A. Avec la topologie produit, Ω est un espace compact sur lequel agit la transformation θ définie par $(\theta\omega)_j = \omega_{j+1}$, appelée *shift-transformation.* Si m désigne la probabilité produit sur Ω, le triplet (Ω, m, θ) est un système dynamique qui joue un rôle important en théorie de l'information. Choisissons la partition :

$$\Pi = \{\Omega_1, ... \Omega_n\}$$

où $\Omega_i = \{\omega | \omega_0 = a_i\}$. Il est clair que $m(\Omega_i) = 1/n$ et que $H(\Pi)$ = log n. On peut alors prouver que l'entropie $\hat{H}$ de ce système dynamique est log n, ce qui montre en particulier que les systèmes obtenus pour des valeurs distinctes de n ne sont pas spatialement isomorphes.

5. Théorie ergodique, probabilités et potentiels

Les problèmes de convergence qui sont abordés au chapitre 2 concernent l'opérateur $T : f \mapsto Tf = f \circ \theta$ agissant dans $L^1(m)$ ou $L^2(m)$. Cet opérateur possède les propriétés qui suivent :

a) T est linéaire ;

b) T est positif : $f \geqslant 0 \Rightarrow Tf \geqslant 0$;

c) T est une contraction, c'est-à-dire : $\| Tf\|_1 \leqslant \|f\|_1$ pour tout $f \in L^1(m)$.

On peut aussi considérer de façon plus générale des endomorphismes de l'espace (réel) $L^1(m)$ possédant les propriétés précédentes et non nécessairement induits par des transformations ponctuelles θ. De tels opérateurs se présentent naturellement dans la théorie des processus markoviens. Ils sont définis à partir d'un noyau N : $\Omega \times \mathcal{B} \to \bar{\mathbf{R}}^+$ ($\mathcal{B}$ désignant la tribu des ensembles mesurables de Ω) où l'on suppose que l'application partielle $\omega \mapsto N(\omega, A)$ est mesurable pour A constant et que l'application $A \mapsto N(\omega, A)$, ω étant fixé, est **une** probabilité (ou une sous-probabilité) sur $\mathcal{B}$. **À** tout $f \in L^1$ on associe la mesure réelle μ_f sur $(\Omega, \mathcal{B})$ par la formule :

$$\mu_f(A) = \int_\Omega N(\omega, A) f(\omega)\, dm(\omega), \quad A \in \mathcal{B}.$$

Si l'on suppose que μ_f est absolument continue par rapport à m, et cela quel que soit f, la densité $Tf = d\mu_f/dm$ est la transformée de f par T. La vérification des propriétés a, b et c est immédiate.

Le théorème ergodique général, établi en 1960 par Chacon et Ornstein, pour une contraction positive T affirme que : Quelles que soient les fonctions intégrables f et g, $g \geqslant 0$, l'expression :

$$\sum_{k=0}^{n} T^k f \Big/ \sum_{k=0}^{n} T^k g$$

tend presque partout vers une limite finie sur l'ensemble :

$$\left\{ \sum_{k=0}^{\infty} T^k g > 0 \right\}.$$

Chacon a, de plus, explicité cette limite. Cet important théorème, faisant suite à des travaux de Doob et de E. Hopf, a été aussi prouvé par J. Neveu par des méthodes probabilistes.

Le lien avec la théorie du potentiel découle de recherches faites par A. Brunel.

par P. A. Meyer et par Ackoglu qui ont utilisé le lemme suivant, appelé *lemme ergodique maximal.*

Soit $f \in L^1$ (réel) et $A \in \mathcal{B}$ tel que :

$$A \subset \left\{ \limsup_n \left(\sum_{k=0}^{n} T^k f \right) > 0 \right\},$$

alors :

$$\int e_A . f \, dm \geqslant 0,$$

en désignant par e_A le potentiel d'équilibre de A, relativement au noyau transposé ${}^tT$ de T.

L'opérateur ${}^tT$ qui agit dans L^∞ se définit par dualité : Quels que soient $f \in L^1$ et $g \in L^\infty$,

$$\int Tf . g \, dm = \int f . {}^tTg \, dm ;$$

la fonction e_A est, en gros, la plus petite fonction ${}^tT$ – sous-invariante (${}^tTe_A \leqslant e_A$) qui majore 1_A.

On a voulu montrer comment la théorie ergodique s'est développée et ramifiée à partir du problème fondamental posé par l'hypothèse ergodique. Il n'était pas possible de résumer ici d'autres travaux difficiles, par exemple ceux qui concernent les groupes ou les semi-groupes de transformations ou d'opérateurs.

ANTOINE BRUNEL

Bibliographie

P. BILLINGSLEY, *Ergodic Theory and Information*, Krieger Publ., New York, 1978 / N. A. FRIEDMANN, *Introduction to Ergodic Theory*, Van Nostrand, New York. 1970 / U. KRENGEL & A. BRUNEL, *Ergodic Theorems, with a Supplement on Harris Processes*, De Gruyter, Hawthborne (N. Y.), 1985 / P. A. MEYER, « Théorie ergodique et potentiels », in *Ann. Inst. Fourier*, t. XV, fasc. I, 1965 / K. PETERSEN, *Ergodic Theory*, Cambridge Univ. Press, 2' éd. 1990 / D. REVUZ, *Markov Chains*, North-Holland, Amsterdam, rééd. 1991.

ESPACE AFFINE → AFFINES ESPACE & REPÈRE

ESPACE DE HILBERT → HILBERT ESPACE DE

ESPACE PROJECTIF → PROJECTIFS ESPACE & REPÈRE

ESPACES MÉTRIQUES → MÉTRIQUES ESPACES

ESPACES VECTORIELS NORMÉS → NORMÉS ESPACES VECTORIELS

ESPACES VECTORIELS TOPOLOGIQUES → TOPOLOGIQUES ESPACES VECTORIELS

EXPONENTIELLE & LOGARITHME

Pour les constructeurs des premières tables, les logarithmes étaient avant tout un outil de calcul numérique ; mais leur importance n'a cessé de croître. De nos jours, les logarithmes et les exponentielles interviennent dans tous les domaines de l'activité humaine, qu'il s'agisse de physique, de médecine, de sciences humaines... C'est le cas de tout phénomène naturel dans lequel deux mesures x et y sont telles que le taux de variation $\Delta y/\Delta x$ de y est proportionnel à y ; la quantité y dépend alors exponentiellement de x, car on a y' = ky. Mais les exponentielles s'introduisent aussi dans de nombreux autres cas ; c'est ainsi que les lois de Laplace-Gauss ou de Poisson sont des techniques de base de la statistique.

En tant que fonctions nouvelles, les transcendantes élémentaires (logarithmes, exponentielles et fonctions trigonométriques) se sont introduites d'une façon naturelle au cours du XVIIe siècle, à partir de considérations cinématiques tout d'abord (étude de la cycloïde par exemple). Avec les débuts du calcul infinitésimal, ces fonctions acquièrent une grande importance théorique : découverte de leurs développements en série et rôle essentiel qu'elles jouent dans l'intégration de nombreuses équations différentielles simples. Au XVIIIe siècle, le mathématicien suisse L. Euler, par extension au champ complexe, a mis en évidence les liens étroits qui existent entre ces fonctions et a introduit les notations que l'on utilise encore aujourd'hui.

Dans ce qui suit, on construit complètement ces fonctions à partir du logarithme népérien, primitive de $1/x$, en se limitant à l'aspect théorique sans aborder l'aspect pratique des calculs.



1. Résultats préliminaires

Soit R le groupe additif des nombres réels ; les nombres réels strictement positifs forment un groupe pour la multiplication que nous noterons RT. On se propose ici de décrire tous les homomorphismes continus de ces groupes entre eux. Ainsi, les fonctions *logarithmes,* les fonctions *exponentielles* et les fonctions *puissances* sont des applications continues f, g, h :

$$f:\mathbf{R}_+^* \to \mathbf{R},\ g:\mathbf{R}\to\mathbf{R}_+^*,\ h:\mathbf{R}_+^*\to\mathbf{R}_+^*,$$

qui vérifient respectivement les relations fonctionnelles :

$$f(xy) = f(x) + f(y), \qquad g(x+y) = g(x)g(y),$$
$$h(xy) = h(x)h(y).$$

Montrons pour commencer que les seuls homomorphismes continus du groupe additif R dans lui-même sont les *homothéties.* Soit donc :

$$u:\mathbf{R}\to\mathbf{R}$$

une application continue telle que $u(x+y) = u(x) + u(y)$; posons $u(1) = a$. Pour montrer que $u(x) = ax$ pour tout nombre réel x, il suffit, à cause de la continuité, d'établir ce résultat pour x rationnel. Remarquons d'abord que la relation fonctionnelle entraîne :

$$u(n) = u(\underbrace{1+1+\dots+1}_{n \text{ fois}}) = \underbrace{a+a+\dots+a}_{n \text{ fois}} = an,$$

pour tout entier positif n ; d'autre part, $u(1) = u(1+0) = u(1) + u(0)$, d'où $u(0) = 0$. Pour les entiers négatifs, on a : $0 = u(0) = u(n + (-n)) = u(n) + u(-n)$, d'où :

$$u(-n) = -u(n) = a \times (-n);$$

ainsi, $u(x) = ax$ pour tout entier relatif. Soit enfin $x = p/q$ un nombre rationnel ; on a :

$$\underbrace{x + x + \dots + x}_{q \text{ fois}} = qx = p,$$

d'où :

$$u(qx) = qu(x) = u(p) = a\,xp,$$

et finalement :

$$u(x) = a \times \frac{p}{4} = ax.$$

Pour $a = 0$, on obtient l'application nulle et, pour $a \neq 0$, ces homomorphismes sont des *isomorphismes,* c'est-à-dire qu'ils sont bijectifs.

Il est facile de voir que la continuité de u équivaut à la continuité à l'origine, ou encore au fait que u soit bornée au voisinage de zéro. On peut même démontrer que la mesurabilité de u suffit; en revanche, si l'on n'impose aucune condition, on peut montrer, en faisant appel à l'axiome du choix, qu'il existe des homomorphismes u autres que les homothéties.

Revenons aux équations fonctionnelles vérifiées par f et g. En intégrant ces équations, on voit que f et g sont, en fait, de classe C^1. On peut donc dériver des équations par rapport à y ; ce qui donne :

$$xf'(xy) = f'(x)$$

et donc, pour $y = 1$,

$$xf'(x) = f'(1);$$

de même, on a :

$$g'(x+y) = g'(y)g(x)$$

et donc, pour $y = 0$,

$$g'(x) = g'(0)g(x).$$

En particulier, on voit que $f'(x) = k/x$, ce qui conduit à étudier les primitives de $x \mapsto 1/x$. C'est ainsi que l'on définira le logarithme au chapitre 2. La fonction exponentielle s'en déduit alors par passage à la fonction réciproque (chap. 3). Nous aurons besoin pour cela du théorème classique d'inversion.

Inversion des fonctions monotones

Dans ce qui suit, nous nous limiterons, pour des facilités d'énoncé, à des fonctions croissantes, étant entendu que les résultats correspondants pour les fonctions décroissantes s'en déduisent immédiatement.

Soit f une fonction à valeurs réelles définie et *strictement croissante* dans un intervalle $I = (a, b)$; I est quelconque, borné ou pas (ce qui veut dire qu'on peut avoir $a = -\infty$ par exemple), ouvert, fermé ou semi-ouvert. Par des arguments très analogues à ceux qui sont exposés à la fin du chapitre 4 de l'article CALCUL INFINITÉSIMAL - Calcul à une variable, on peut montrer que $f(x)$ tend vers une limite α (resp. β), éventuellement égale à $-\infty$ (resp. $+\infty$), pour x tendant vers a par valeurs supérieures (resp. vers b par valeurs inférieures) : cette limite en a est la borne inférieure (ou $-\infty$, si cet ensemble n'est pas borné inférieurement) de l'ensemble des nombres $f(x)$, $x \in I$.

Si f est de plus continue, on déduit facilement du théorème 14 *bis* du même article (chap. 9), dit théorème des valeurs intermédiaires, que f réalise une bijection

de 1 sur l'intervalle J = (α, β), les extrémités correspondantes de I et J étant de même nature, incluses ou exclues simultanément. De plus, la bijection réciproque de J sur 1 est, sous ces hypothèses, continue. Ainsi, on peut énoncer le résultat suivant : si f est une application continue strictement croissante d'un intervalle I dans R, alors c'est une bijection de I sur l'intervalle image et la bijection réciproque est continue. Si f est dérivable, l'application réciproque est dérivable en tout point $x = f(y) \in J$ tel que $f'(y) \neq 0$ et on a :

$$(f^{-1})'(x) = \frac{1}{f'(y)}.$$

On verra dans ce qui suit de nombreux exemples de cette situation.

2. Logarithmes

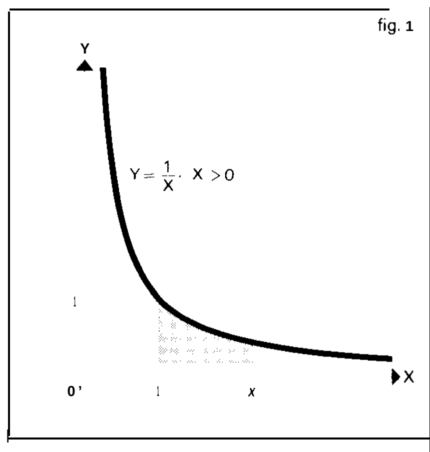
Définition

Il n'existe pas de fonction rationnelle admettant pour dérivée $1/x$; pourtant, cette fonction est définie et continue pour $x > 0$, et, par suite (cf. CALCUL INFINITÉSIMAL Calcul à une variable, chap. 5), elle admet des primitives dans cet intervalle. Ces primitives constituent donc de « nouvelles » fonctions dont nous allons étudier les propriétés. Elles diffèrent toutes entre elles d'une constante, et il suffit d'en examiner une.

On appelle *logurithme népérien ou naturel* la primitive de $1/x$ dans $]0, \infty[$ qui s'annule pour $x = 1$, soit :

$$(1) \qquad L(x) = \ln x = \int_1^x \frac{dt}{t}, \quad x > 0;$$

ainsi, c'est une fonction dérivable, de dérivée $1/x$. Géométriquement, si $x > 1$, c'est la mesure de l'aire comprise entre l'hyperbole d'équation $Y = 1/X$ et les deux droites $X = 1$ et $X = x$ (fig. 1) ; on a donc $\ln x < 0$ pour $0 < x < 1$ et $\ln x > 0$ pour $x > 1$. On utilisera dans l'ouvrage la notation normalisée anglo-saxonne $\ln x$



Interprétation géométrique du logarithme népérien

Établissons dès maintenant la propriété fondamentale du logarithme népérien : c'est un homomorphisme (en fait, comme on le verra ci-dessous, un *isomorphisme)* du groupe multiplicatif $\mathbf{R}_+^*$ dans le groupe additif R. Soit y un nombre réel positif; la fonction $f(x) = \ln xy$ a la même dérivée que la fonction $\ln x$:

$$f'(x) = y \, . \, L'(xy) = y \frac{1}{xy} = \frac{1}{x},$$

et, par suite, ces deux fonctions diffèrent d'une constante, soit :

$$\ln xy = \ln x + k;$$

faisant $x = 1$, on a $k = \ln y$, d'où la relation fonctionnelle :

$$(2) \qquad \ln xy = \ln x + \ln y.$$

On en déduit immédiatement, pour tout entier $n \in Z$,

$$\ln x^n = n \ln x;$$

plus généralement, avec la convention des

exposants fractionnaires ; si $a = p/q$, $q > 0$, rappelons que, par définition, $x^{\alpha} = \sqrt[q]{x^p}$; on a donc :

(3) $$\ln x^{\alpha} = \ln \sqrt[q]{x^p} = \frac{1}{q} \ln x^p = \frac{p}{4} \ln x = \alpha \ln x.$$

D'autre part, si x et y sont des nombres positifs quelconques, on a $y(x/y) = x$, d'où :

(4) $$\ln \frac{x}{y} = \ln x - \ln y.$$

Comportement et graphe

La fonction logarithme népérien est *strictement croissante* pour $x > 0$, car sa dérivée est strictement positive dans cet intervalle.

Étudions le comportement du logarithme lorsque x tend vers l'infini. Pour tout entier n, on a :

$$\ln 2^n = n \ln 2 ;$$

si A est un nombre positif, soit N un entier plus grand que $A/(\ln 2)$. Pour $x > 2^N = B$, on a :

$$\ln x > \ln 2^N > N \ln 2 > A,$$

ce qui montre que :

(5) $$\ln x \to +\infty, \quad x \to \infty.$$

On en déduit facilement le comportement de $\ln x$ pour x tendant vers 0 par valeurs positives ; si A est un réel positif, on a, pour le même choix de N que ci-dessus,

$$\ln x < -A, \quad \text{pour } x < 1/2^N ;$$

ainsi :

(6) $$\lim_{x \to 0^+} \ln x = -\infty.$$

Précisons le comportement de ln x en montrant que cette quantité est asymptotiquement négligeable devant x pour x tendant vers l'infini. En effet, pour $t \geqslant 1$, on a par exemple :

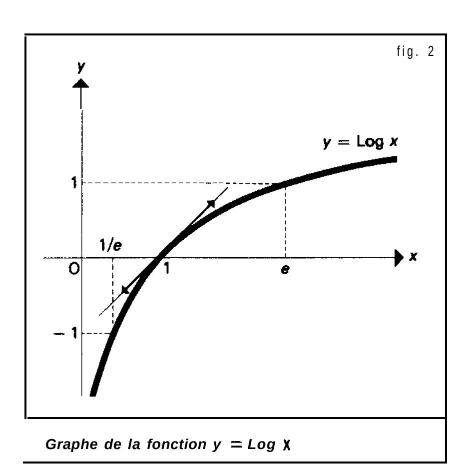
$$\frac{1}{t} \leqslant \frac{1}{\sqrt{t}},$$

d'où, pour $x \geqslant 1$,

$$\ln x = \int_1^x \frac{dt}{t} \leqslant \int_1^x \frac{dt}{\sqrt{t}} = 2(\sqrt{x} - 1) < 2\sqrt{x} ;$$

ainsi : $\ln x/x < 2/\sqrt{x}$, ce qui entraîne bien :

(7) $$\lim_{x \to \infty} \frac{\ln x}{x} = 0.$$



Graphe de la fonction $y = \operatorname{Log} x$

Toutes ces propriétés permettent de tracer le graphe de L (fig. 2). On peut préciser le tracé en remarquant que la fonction est *concave,* car sa dérivée seconde $-1/x^2$ est négative ; la tangente au point d'abcisse 1 est de pente égale à 1, ce qui équivaut à :

(8) $$\lim_{x \to 0} \frac{\ln(1+x)}{x} = 1 ;$$

on exprime cela en disant que $\ln(1+x)$ est équivalent à x pour x tendant vers 0 (cf. calculs ASYMPTOTIQUES). On peut préciser le comportement de $\ln(1+x)$ au voisi-

nage de $x = 0$ par le développement en série :

(9) $\ln(1 + x) = x - x^2/2 + x^3/3 - \ldots + (-1)^{n+1}x^n/n + \ldots$

valable pour $|x| < 1$, qui s'obtient en intégrant terme à terme la série géométrique :

$$1 - x + x^2 - x^3 + \ldots + (-1)^n x^n + \ldots$$

dont la somme est égale à la dérivée $1/(1 + x)$ de $\ln(1 + x)$.

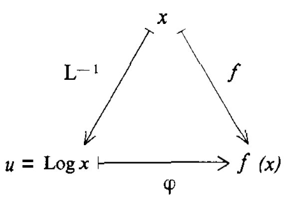
C'est à partir de cette série que l'on peut calculer les valeurs numériques des logarithmes. En définitive, le logarithme népérien $L : \mathbf{R}_+^* \to \mathbf{R}$ est continu, strictement croissant et tend vers $-\infty$ et $+\infty$ pour x tendant vers 0 (par valeurs supérieures) et vers $+\infty$ respectivement ; d'après le théorème d'inversion (cf. chap. 1) c'est donc une bijection, c'est-à-dire un isomorphisme (continu) du groupe multiplicatif des nombres réels positifs sur le groupe additif de tous les nombres réels. La bijection réciproque est un isomorphisme du groupe R sur le groupe RT : c'est la fonction exponentielle que nous examinerons dans le chapitre suivant. En particulier. il existe un unique nombre réel positif dont le logarithme népérien est égal à 1 ; c'est le célèbre nombre e, dont la transcendance a été établie par C. Hermite en 1873.

Autres logarithmes

Proposons-nous maintenant de caractériser tous les homomorphismes continus du groupe multiplicatif RT dans le groupe additif R. Soit f l'un d'entre eux ; tout nombre réel u s'écrit de manière unique :

$$u = \ln x, \quad x > 0,$$

avec $x = L^{-1}(u)$; l'application $\varphi : \mathbf{R} \to \mathbf{R}$, qui, à $u \in \mathbf{R}$, fait correspondre $f(x) = f(L^{-1}(u))$, en accord avec le diagramme suivant :



est un homomorphisme continu du groupe additif R dans lui-même et est donc (cf. chap. 1) de la forme $\varphi(u) = ku$, où u est une constante. Ainsi $f(x) = k \ln x$, c'est-à-dire que f est un multiple scalaire du logarithme népérien.

Soit $f = k\,L$ un tel logarithme ; il existe un unique nombre $a > 0$, défini par $\ln a = 1/k$, tel que $f(a) = 1$. On dit que c'est la hase de la fonction logarithme f et on note :

$$f(x) = \log_a x ;$$

avec cette définition, le logarithme népérien est de base e. On a des formules de changement de base du type :

(10) $$\log_a x = \frac{\ln x}{\ln a},$$

Pour les calculs pratiques, on utilise des tables de logarithmes décimaux, de base $a = 10$; le passage d'un système à l'autre s'effectue au moyen des formules :

$$\log_{10} x = M \ln x, \quad \ln x = \frac{1}{M}\log_{10} x,$$

avec : $M = \dfrac{1}{\ln 10} \sim 0{,}434\ 294\ 481$

et : $\dfrac{1}{M} \sim 2{,}302\ 585\ 093.$

Historique

L'idée qui est historiquement à la base de la notion de logarithme est la comparaison

de la suite des entiers et de la suite des puissances correspondantes d'un nombre a :

$$1, 2, 3, \ldots. n$$
$$a, a^2, a^3, \ldots, a^n,$$

déjà étudiée par Archimède dans son traité de l'*Arénaire* ; dans le *Triparty en la science des nombres* (1484) Nicolas Chuquet remarqua que, si on fait correspondre les termes de même rang, à la somme de deux nombres de la progression arithmétique correspond le produit des nombres de la progression géométrique. Cette correspondance fut étendue aux exposants négatifs et fractionnaires par Michael Stifel (1544), mais la notion de logarithme ne se développa vraiment qu'au début du XVIIe siècle, lorsque l'Écossais John Napier, ou Neper (*Mirifici logarithmorum cunonis descriptio*, Édimbourg, 1614), puis le Suisse Joost Bürgi (*Aritmetische und geometrische Progresstabulen*, Prague, 1620) eurent l'idée d'introduire des nombres intercalaires en quantité suffisante et construisirent des tables permettant de passer d'une progression à l'autre. Neper a rendu « continue » la correspondance entre les deux progressions en utilisant une image cinématique ; il les supposa engendrées l'une et l'autre par mouvement continu (« par fluxion ») : deux points mobiles se déplacent le long d'une droite à partir d'une même position initiale, l'un M avec une vitesse uniforme, l'autre N avec une vitesse proportionnelle à son abscisse ; le logarithme de l'abscisse de N est alors par définition l'abscisse de M.

Les premières tables de logarithmes décimaux sont dues à Henry Briggs (*Arithmetica logarithmica*, 1624) qui fait des logarithmes un moyen de calcul numérique pratique. Les fonctions logarithme et exponentielle s'introduisent en analyse au cours du XVIIe siècle et sont intimement liées à la création et au développement du calcul infinitésimal.

3. Exponentielles réelles

On va maintenant définir la fonction exponentielle comme fonction réciproque du logarithme népérien.

La fonction exponentielle

On appelle *fonction exponentielle* l'isomorphisme E : R $\rightarrow \mathbf{R}_+^*$, réciproque du logarithme népérien ; ainsi, pour tout nombre réel x, E(x) = exp x est l'unique nombre réel > 0 dont le logarithme népérien est égal à x, soit :

(11) $y = \exp x \Leftrightarrow x = \ln y,\ y > 0$;

cela entraîne aussi, par composition de L et E que, pour tout x $\in$ R et pour tout $y \in \mathbf{R}_+^*$, on a:

(12) $\exp(\ln y) = y, \quad \ln(\exp x) = x.$

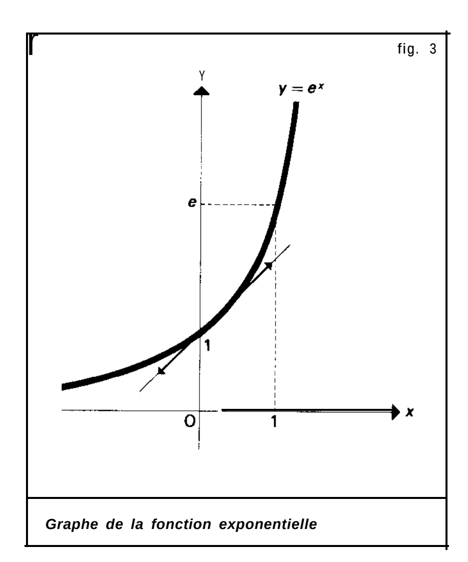
Puisque la fonction logarithme népérien est strictement croissante et dérivable de dérivée toujours non nulle, il en est de même de la fonction exponentielle ; son graphe est le symétrique du graphe de L par rapport à la première bissectrice d'équation y = x (fig. 3).

La dérivée en x de la fonction exponentielle est l'inverse de la dérivée de la fonction logarithme népérien au point $y = \exp x$ soit :

(13) $E'(x) = \dfrac{1}{L'(y)} = y = E(x).$

Plus précisément, la fonction exponentielle est l'unique solution sur R du problème de Cauchy :

(13′) $y' = y,\ y(0) = 1$;



Graphe de la fonction exponentielle

ainsi la fonction exponentielle est indéfiniment dérivable et égale à toutes ses dérivées. La formule de Taylor en 0 s'écrit ici (cf. formule 47, CALCUL INFINITÉSIMAL-Calcul à une variable) :

$$E(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + + \frac{x^n}{n!} + R_n,$$

où : $$|R_n| \leqslant \exp|x| \frac{|x|^{n+1}}{(n+1)!},$$

qui, pour tout x, tend vers 0 lorsque n tend vers l'infini. En effet, pour tout x, la suite $u_p = x^p/p!$ vérifie $u_{p+1}/u_p = x/(p+1)$, qui est inférieur à $1/2$ en valeur absolue pour p assez grand, disons $p \geqslant P$; pour $p \geqslant P$, on a donc $u_{p+1} \leqslant (1/2)^{p-P}$, ce qui montre que la suite u_p tend vers 0 pour p tendant vers $+\infty$. Ainsi, E(x) est, pour tout $x \in$ R, la somme de sa série de Taylor, soit :

(14) $$\exp x = 1 + x + x^2/2! + \dots + x^n/n! + \dots;$$

Pour $x \to +\infty$, on déduit facilement de (7) le comportement asymptotique :

(15) $$\lim_{x \to \infty} \frac{x^n}{\exp x} = 0, n \in \mathbf{N}.$$

Soit enfin une dernière propriété de la fonction exponentielle. Pour tout entier n :

$$\exp x = \left(\exp \frac{x}{n}\right)^n;$$

puisque :

$$\exp \frac{x}{n} = 1 + \frac{x}{n} + o\left(\frac{1}{n}\right),$$

cela conduit à la formule :

$$\exp x = \lim_{x \to \infty}\left(1 + \frac{x}{n}\right)^n;$$

ce type de raisonnement (dû en subtance à Euler) demande, bien entendu, à être établi rigoureusement, par exemple par les développements limités.

Le nombre e

Pour x = 1, E(1) = e, base des logarithmes népériens. Ce nombre est la somme de la série :

$$1 + 1 + 1/2 + 1/6 + \dots + 1/n! + \dots;$$

c'est aussi la limite de l'expression :

$$\left(1 + \frac{1}{n}\right)^n$$

pour n tendant vers l'infini.

Une valeur approchée de e, à 10^{-24} près, est :

2,718 281 828 459 045 235 360 287.

Si n est un entier relatif, on a E(n) = e^n, puisque ln e^n = n ln e = n. Plus généralement, si $x = p/q$, $q > 0$, est un nombre rationnel, on a :

$$E(x) = \sqrt[q]{e^p} = e^{p/q} = e^x,$$

avec la convention des exposants fractionnaires. Ainsi, la fonction exponentielle est un prolongement continu à R tout entier de l'application $x \mapsto e^x$ de Q dans RT définie par la convention des exposants

fractionnaires ; cela conduit à généraliser cette notation en posant *pur définition :*

$$E(x) = e^x,$$

pour tout réel x ; le fait que l'exponentielle est un homomorphisme de groupe s'écrit maintenant :

$$(16) \qquad e^{x+y} = e^x e^y,$$

formule valable pour x et y réels quelconques.

Trigonométrie hyperbolique

Introduisons maintenant les fonctions hyperboliques, qui jouent pour la géométrie du plan hyperbolique le même rôle que les fonctions circulaires pour le plan euclidien (cf. GROUPES - Groupes classiques et géométrie, chap. 3).

Pour tout nombre réel x, on appelle *cosinus hyperbolique* de x, *sinus hyperbolique* de x et *tangente hyperbolique* de x respectivement les nombres :

$$(17) \qquad \begin{cases} \operatorname{ch} x = \dfrac{1}{2}(e^x + e^{-x}) \\ \operatorname{sh} x = \dfrac{1}{2}(e^x - e^{-x}) \\ \operatorname{th} x = \dfrac{\operatorname{sh} x}{\operatorname{ch} x} = \dfrac{e^{2x} - 1}{e^{2x} + 1}; \end{cases}$$

remarquons que le cosinus hyperbolique est une fonction paire, tandis que les deux autres sont des fonctions impaires.

Un calcul simple montre que l'on a :

$$\operatorname{ch}^2 x - \operatorname{sh}^2 x = 1$$

pour tout x ; si a et b sont deux nombres réels la propriété (16) de la fonction exponentielle entraîne les formules d'addition suivantes :

$$\operatorname{ch}(a+b) = \operatorname{ch} a \operatorname{ch} b + \operatorname{sh} a \operatorname{sh} b,$$
$$\operatorname{sh}(a+b) = \operatorname{sh} a \operatorname{ch} b + \operatorname{ch} a \operatorname{sh} b.$$

Par dérivation dans (17), on obtient facilement :

$$(\operatorname{ch} x)' = \operatorname{sh} x, \quad (\operatorname{sh} x)' = \operatorname{ch} x,$$
$$(\operatorname{th} x)' = 1 - \operatorname{th}^2 x = \frac{1}{\operatorname{ch}^2 x}.$$

Puisque $\operatorname{ch} x \geqslant 1$, il en résulte que $\operatorname{sh} x$ est une application strictement croissante de R dans R et on voit que $\operatorname{sh} x$ tend vers $-\infty$ et $+\infty$ pour x tendant vers $-\infty$ et $+\infty$ respectivement : cette fonction réalise donc une bijection de R sur R. Puisque $\operatorname{sh} 0 = 0$, le nombre $\operatorname{sh} x$ est du signe de x ; par suite la fonction paire $\operatorname{ch} x$ est strictement décroissante pour $x \leqslant 0$ et strictement croissante pour $x \geqslant 0$; d'autre part, $\operatorname{ch} x$ tend vers $+\infty$, pour x tendant vers $+\infty$ ou vers $-\infty$. Enfin, la fonction $\operatorname{th} x$, dont la dérivée est toujours > 0, est strictement croissante ; pour x tendant vers l'infini, on a :

$$\lim_{x \to -\infty} \operatorname{th} x = -1, \quad \lim_{x \to +\infty} \operatorname{th} x = +1,$$

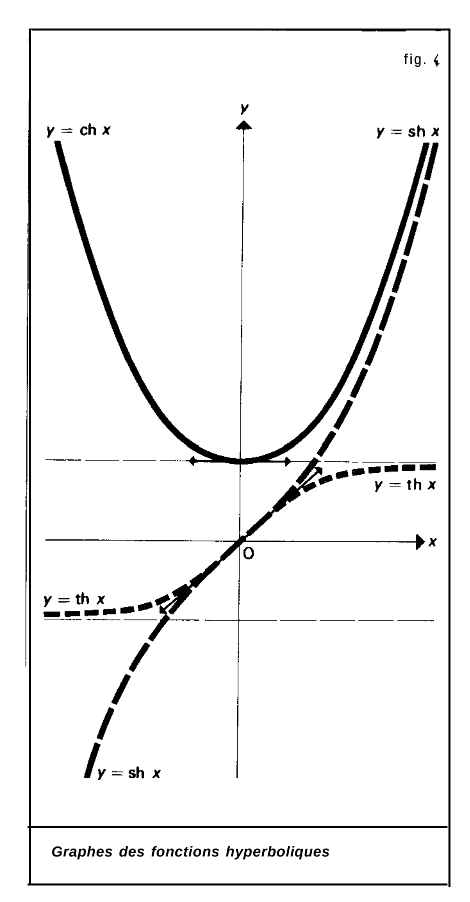
et, par suite, $\operatorname{th} x$ réalise une bijection de R sur l'intervalle $]-1, +1[$.

On a représenté les graphes des fonctions hyperboliques sur la figure 4.

L'application $t \mapsto (\operatorname{ch} t, \operatorname{sh} t)$ est une représentation paramétrique de la branche d'hyperbole $x^2 - y^2 = 1$, $x > 0$, d'où le nom de *trigonométrie hyperbolique,* par analogie avec la trigonométrie « circulaire ». La bijection réciproque s'appelle *amplitude* et se note $u \mapsto \operatorname{Am} u$.

On désigne par $\operatorname{Arg\,sh} x$ la bijection de R sur R réciproque de $\operatorname{sh} x$, par $\operatorname{Arg\,ch} x$ la bijection de $]1, \infty[$ sur $\mathbf{R}_+$ réciproque de la restriction de ch x à $\mathbf{R}_+$ et enfin par $\operatorname{Arg\,th} x$ la bijection de $]-1, +1[$ sur R réciproque de $\operatorname{th} x$. Ainsi :

$$y = \operatorname{Arg\,sh} x, \quad x \in \mathbf{R} \Leftrightarrow x = \operatorname{sh} y, \quad y \in \mathbf{R};$$
$$y = \operatorname{Arg\,ch} x, \quad x \geqslant 1 \Leftrightarrow x = \operatorname{ch} y, \quad y \geqslant 0;$$
$$y = \operatorname{Arg\,th} x, \quad -1 < x < +1 \Leftrightarrow x = \operatorname{th} y, \quad y \in \mathbf{R}.$$



Graphes des fonctions hyperboliques

On obtient $e^{2y} - 2xe^y - 1 = 0$, en remplaçant sh y par son expression exponentielle dans $x = \text{sh}\, y$; d'où $e^y = x + \sqrt{x^2 + 1}$, puisque $e^y > 0$. Ainsi, Arg shx s'exprime au moyen du logarithme népérien par la formule :

$$\text{Arg sh}\, x = \ln(x + \sqrt{x^2+1}) ;$$

par des raisonnements analogues, on aurait :

$$\text{Arg ch}\, x = \ln(x + \sqrt{x^2-1}),$$

$$\text{Arg th}\, x = \frac{1}{2} \ln \frac{1+x}{1-x}$$

Les dérivées des fonctions hyperboliques inverses sont très simples et se calculent en appliquant le théorème du chapitre 1 sur la dérivée d'une fonction réciproque ; nous ne détaillerons pas le calcul, nous contentant d'indiquer le résultat :

$$(\text{Arg ch}\, x)' = \frac{1}{\sqrt{x^2-1}},$$

$$(\text{Arg sh}\, x)' = \frac{1}{\sqrt{x^2+1}},$$

$$(\text{Arg th}\, x)' = \frac{1}{1-x^2}.$$

Pour terminer ces quelques lignes sur la trigonométrie hyperbolique, donnons enfin les développements en série de ch x, sh x, qui se déduisent immédiatement de (14) et de (17), et de Arg th x, qui s'obtient par intégration à partir de la série géométrique :

$$\frac{1}{1-x^2} = 1 + x^2 + x^4 + \dots.$$

avec $|x| < 1$:

$$\text{ch}\, x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \dots + \frac{x^{2n}}{(2n)!} + \dots,\ x \in \mathbf{R}$$

$$\text{sh}\, x = x + \frac{x^3}{3!} + \dots + \frac{x^{2n+1}}{(2n+1)!} + \dots,\ x \in \mathbf{R} ;$$

$$\text{Arg th}\, x = x + \frac{x^3}{3} + \frac{x^5}{5} + \dots + \frac{x^{2n+1}}{2n+1} + \dots,$$
$$-1 < x < +1.$$

Fonction exponentielle de base a

Soit a un nombre réel strictement positif. Pour x rationnel, on a $\ln a^x = x \ln a$ (avec la convention des exposants fractionnaires ; cf. *supra.* chap. 1) ; d'où, d'après (12):

$$(18) \qquad a^x = e^{x \ln a}.$$

Désirant étendre cette notation, nous prendrons l'expression (18) comme *définition* de a^x pour tout x réel ; la fonction E_a, définie par $E_a(x) = a^x$ s'appelle la fonction exponentielle de base a.

Pour $a = 1$, on a $1^x = 1$ pour tout nombre réel x. Pour a > 0, $a \neq 1$, faisons le lien avec le logarithme de base a ; le logarithme de base a réalise un isomorphisme (strictement croissant si $a > 1$. strictement décroissant si $a < 1$) du groupe multiplicatif R* sur le groupe additif R ; l'isomorphisme réciproque fait correspondre à tout réel x l'unique nombre $y > 0$ tel que :

$$x = \log_a y ;$$

puisque :

$$x = \log_a y = \frac{\ln y}{\ln a},$$

d'où $\ln y = x \ln a$; on en déduit :

$$y = e^{x \ln a} = a^x,$$

ce qui montre que la fonction exponentielle de base a est la fonction réciproque du logarithme de base *a.* C'est donc un isomorphisme de groupes, d'où :

(19) $$a^0 = 1, \quad a^{-x} = 1/a^x, \quad a^{x+y} = a^x a^y.$$

Les fonctions $x \mapsto a^x$ sont les seuls homomorphismes continus du groupe additif R dans le groupe multiplicatif $\mathbf{R}_+^*$.

De la définition (18) résulte aussi :

$$(a^x)^y = a^{xy}, \quad (ab)^x = a^x b^x,$$

pour a, b réels positifs et x, y réels quelconques. En effet, pour la première de ces formules par exemple, on a :

$$(a^x)^y = e^{y \ln a^x} = e^{yx \ln a} = a^{xy}$$

La fonction exponentielle de base a est strictement croissante pour $a > 1$; elle est strictement décroissante pour $a < 1$; sa dérivée s'obtient facilement sous la forme (18) ; on a :

$$(a^x)' = (e^{x \ln a})' = (\ln a) e^{x \ln a} = (\ln a) a^x.$$

Fonctions puissances

Soit u un nombre réel quelconque ; on appelle *fonction puissance* l'application P_u de R* dans $\mathbf{R}_+^*$ qui, à $x > 0$, fait correspondre le nombre réel positif x^u. Par définition, on a donc :

$$P_u(x) = x^u = e^{u \ln x} :$$

de (19) résulte que l'on a, pour $x > 0$, $y > 0$,

$$P_u(xy) = P_u(x) P_u(y),$$

c'est-a-dire que P_u est un homomorphisme du groupe multiplicatif $\mathbf{R}_+^*$ dans lui-même ; il résulte de ce qui précède que c'est un isomorphisme pour $u \neq 0$, l'isomorphisme réciproque étant $P_u^{-1} = P_{1/u}$. Pour $u \geqslant 0$, P_u est strictement croissant et se prolonge par continuité pour $x = 0$ en posant $P_u(0) = 0$; pour $u < 0$, P_u est strictement décroissant. Les fonctions $x \mapsto x^u$ sont les seuls homomorphismes continus de $\mathbf{R}_+^*$ dans $\mathbf{R}_+^*$. Dans tous les cas, on a, pour $x > 0$,

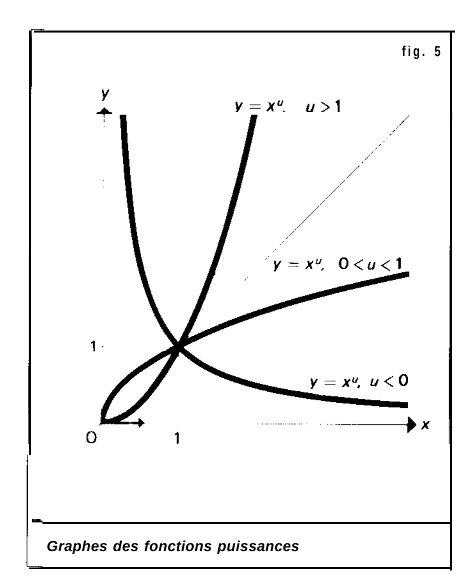
$$P'_u(x) = uP_{u-1}(x) = ux^{u-1} ;$$

sur la figure 5. on a représenté les graphes des fonctions puissances pour diverses valeurs de u.

Indiquons que l'on peut montrer. par des majorations explicites que nous ne ferons pas ici, que le reste de la formule de Taylor de P_u au voisinage de 1 tend vers 0, pour x assez voisin de 1 ; on obtient ainsi le développement en série :

$$(1+x)^u = 1 + ux + \frac{u(u-1)}{2!}x^2 + \frac{u(u-1)(u-2)}{3!}x^3 + \dots + \frac{u(u-1)\dots(u-k+1)}{k!}x^k + \dots,$$

valable pour $|x| < 1$, qui généralise au cas



Graphes des fonctions puissances

d'un exposant réel quelconque la classique formule du binôme de Newton.

Enfin, décrivons le comportement asymptotique des fonctions puissances pour $u > 0$, la fonction P_u tend vers l'infini « plus vite » que la fonction logarithme et « moins vite » que la fonction exponentielle, lorsque x tend vers l'infini, soit :

$$(20) \qquad \lim_{x \to \infty} \frac{x^u}{a^x} = 0 \quad (u \in \mathbf{R},\ a > 1),$$

$$(21) \qquad \lim_{x \to \infty} \frac{\ln x}{x^u} = 0 \quad (u > 0),$$

comme cela résulte facilement de (15).

4. Extension du domaine complexe et trigonométrie

Nous commencerons par le cas de la fonction exponentielle, le plus simple car le développement en série entière (14) converge encore pour tout x *complexe,* et cela suggère d'étendre cette fonction au domaine complexe en la *définissant*, dans ce cas, comme somme de la série correspondante.

L'exponentielle complexe

La série :

$$(22) \quad 1 + \frac{z}{1} + \frac{z^2}{2} + \frac{z^3}{6} + \dots + \frac{z^n}{n!} \dots = \sum_{n=0} \frac{z^n}{n!}$$

est absolument convergente pour tout nombre complexe z. Pour z réel, la somme est e^z. Pour $z \in \mathbf{C}$, nous noterons encore $\exp z$ ou e^z la somme de cette série. D'après la règle de multiplication des séries absolument convergentes, on a, pour $a, b \in \mathbf{C}$,

$$e^a e^b = \sum_{n=0}^{\infty} \frac{a^n}{n!} \sum_{m=0}^{\infty} \frac{b^m}{m!} = \sum_{n=0}^{\infty} c_n,$$

où :

$$c_n = \sum_{k=0}^{n} \frac{a^k b^{n-k}}{k!\,(n-k)!}$$

$$= \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!} a^k b^{n-k} = \frac{(a+b)^n}{n!};$$

ainsi, on a l'importante formule d'addition :

$$(23) \qquad e^{a+b} = e^a e^b, \quad a, b \in \mathbf{C}$$

Puisque $1 = e^0 = e^z e^{-z}$ pour tout $z \in \mathbf{C}$, on a toujours $e^z \neq 0$ et, ainsi, la formule d'addition exprime que l'exponentielle complexe définit un homomorphisme du groupe additif $\mathbf{C}$ de tous les nombres complexes dans le groupe multiplicatif $\mathbf{C}^*$ des nombres complexes non nuls.

En outre, pour tout $a \in \mathbf{C}$, on a :

$$\exp \bar{a} = \overline{\exp a}.$$

Enfin, par dérivation terme à terme de la série correspondante, on voit que, pour tout nombre complexe a, la fonction de variable réelle

$$\varphi_a : t \mapsto \exp at$$

satisfait à la relation $\varphi_a'(t) = a\varphi_a(t)$. Plus

précisément φ_a est l'unique solution sur R du problème de Cauchy : $y' = ay$, $y(0) = 1$.

Fonctions circulaires

Soit $z = x + iy$ un nombre complexe, on a :

(24) $$e^{x+iy} = e^x e^{iy} ;$$

la fonction exponentielle $e \mapsto e^x$ ayant été étudiée, nous allons examiner maintenant la fonction $y \mapsto e^{iy}$.

Pour t réel, on appelle respectivement *cosinus* et *sinus* de t les parties réelle et imaginaire de e^{it}, soit, par définition,

(25) $$e^{it} = \cos t + i \sin t ;$$

il en résulte immédiatement les « formules d'Euler » :

(26) $$\cos t = \frac{e^{it} + e^{-it}}{2}, \quad \sin t = \frac{e^{it} - e^{-it}}{2i}$$

D'après ce qui précède, l'application $\varphi : t \mapsto \exp it$ est un homomorphisme du groupe additif R dans le groupe multiplicatif U des nombres complexes de module 1 (cf. nombres COMPLEXES) et $\varphi'(t) = i\varphi(t)$. L'étude de ce morphisme constitue ce qu'on appelle traditionnellement la *trigonométrie (fig. 6).*

La relation $e^{it} = 1$ signifie que :

(27) $$\cos^2 t + \sin^2 t = 1,$$

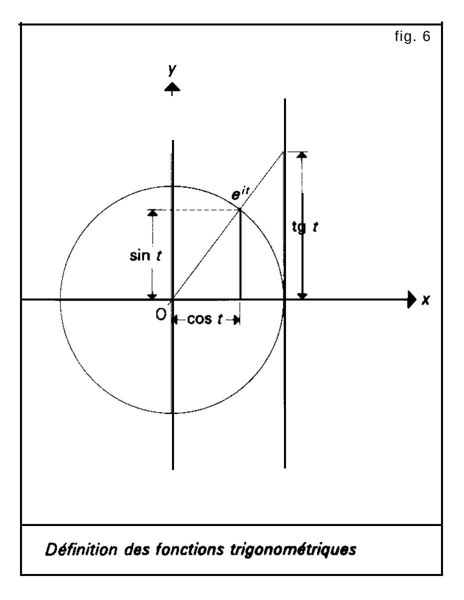
qui est la relation fondamentale de la trigonométrie.

Par ailleurs, la propriété fonctionnelle :

$$e^{i(t+t')} = e^{it} e^{it'}$$

donne, en séparant parties réelle et imaginaire, les formules d'addition de la trigonométrie :

$$\cos(t + t') = \cos t \cos t' - \sin t \sin t',$$
$$\sin(t + t') = \sin t \cos t' + \sin t' \cos t.$$



Définition des fonctions trigonométriques

Remplaçant e^{it} et e^{-it} dans les formules d'Euler par leurs développements en série déduits de (22), on obtient les développements en séries entières, valables pour tout nombre réel t, des fonctions trigonométriques :

(28) $$\cos t = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \ldots + (-1)^n \frac{t^{2n}}{(2n)!} + \ldots,$$

(29) $$\sin t = t - \frac{t^3}{3!} + \ldots + (-1)^n \frac{t^{2n+1}}{(2n+1)!} + \ldots ;$$

ainsi les fonctions sinus et cosinus sont indéfiniment dérivables. Par dérivation des formules d'Euler, ou des développements en série qui précèdent, on a :

(30) $$\cos' = -\sin, \quad \sin' = \cos.$$

Le nombre π

Pour $t = 2$, on a :

$$\cos 2 = 1 - \frac{2^2}{2} + \sum_{n=2}^{\infty} (-1)^n \frac{2^{2n}}{(2n)!}$$

$$< -1 + \sum_{n=2}^{\infty} \frac{2^{2n}}{(2n)!}$$

$$< -1 + \frac{2^4}{4!} \sum_{k=0}^{\infty} \left(\frac{2}{5}\right)^{2k} = -1 + \frac{50}{63} < 0 ;$$

puisque la fonction cosinus est continue et égale à 1 pour $t = 0$, il existe un plus petit nombre réel $\tau > 0$ tel que $\cos \tau = 0$. Nous désignerons par la lettre grecque π, notation traditionnelle depuis Euler, le nombre $\pi = 2\tau$. Ce nombre π, dont la transcendance a été établie par F.Lindemann en 1882, est égal à la moitié de la longueur du cercle de rayon 1. Une valeur approchée à 10^{-20} près est :

$$\pi \sim 3{,}141\ 592\ 653\ 589\ 793\ 238\ 46.$$

Ainsi, par définition de π, on a $\cos t > 0$ dans l'intervalle $]0, \pi/2[$, ce qui entraîne, d'après (30), que la fonction sinus est strictement croissante dans l'intervalle $[0, \pi/2]$. Puisque $\sin 0 = 0$, cette fonction est donc strictement positive dans l'intervalle $]0, \pi/2]$, ce qui entraîne toujours d'après (30), que le cosinus est strictement décroissant dans cet intervalle. On peut alors constituer, entre 0 et $\pi/2$, le tableau de variation des fonctions circulaires.

Pour $t = \pi/2$, la relation (27) entraîne que le sinus est, en valeur absolue, égal à 1 ; par suite puisque ce nombre est positif $\sin \pi/2 = 1$. Ainsi :

$$(31) \qquad e^{i\pi/2} = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2} = i,$$

d'où, en utilisant la formule d'addition,

$$e^{i\pi} = i^2 = -1, \quad e^{2i\pi} = i^4 = 1.$$

Pour tout nombre complexe z, on a donc :

$$(32) \qquad e^{z+2i\pi} = e^z e^{2i\pi} = e^z ;$$

ce qui montre que la fonction exponentielle complexe est périodique, de période imaginaire pure $2\,i\pi$; en particulier, les fonctions circulaires sont périodiques de période 2π et il suffit de les étudier dans l'intervalle $[0, 2\pi]$ par exemple. Leur variation dans cet intervalle se déduit immédiatement de leur variation dans $[0, \pi/2]$ en utilisant les relations :

$$\cos\left(t + \frac{\pi}{2}\right) = -\sin t, \quad \sin\left(t + \frac{\pi}{2}\right) = \cos t,$$

qui ne font qu'exprimer que :

$$e^{i(t+\pi/2)} = e^{it} e^{i\pi/2} = ie^{it} = -\sin t + i \cos t$$

On peut ainsi former le tableau de variation de ces fonctions et construire leurs graphes (fig. 7).

On introduit aussi la *fonction tangente,* définie pour $t \neq (2k + 1)\,\pi/2$, $k \in \mathbf{Z}$ par :

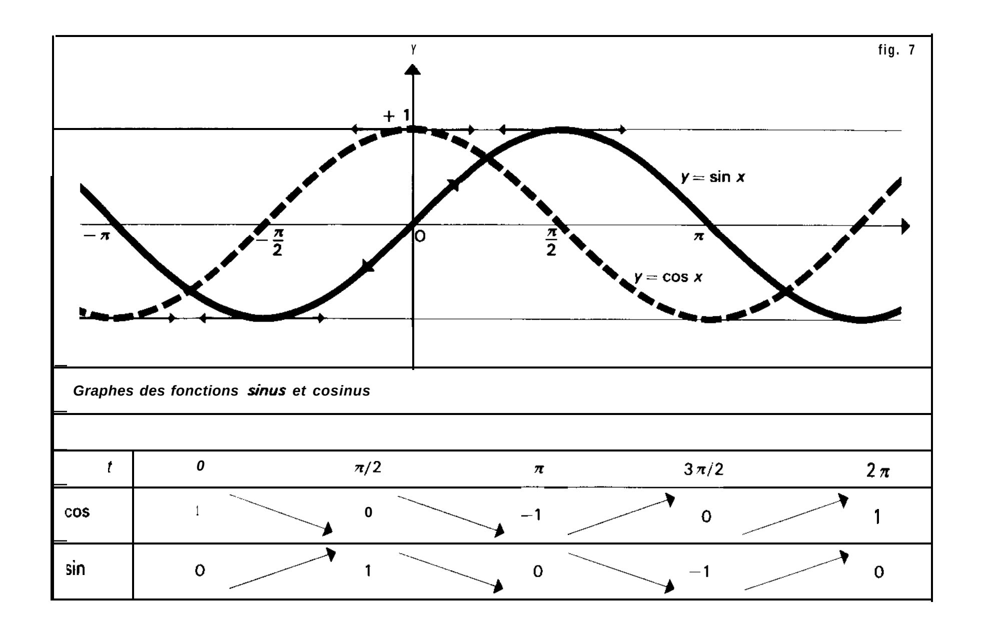
$$\tan t = \frac{\sin t}{\cos t} ;$$

elle est périodique de période π, de dérivée $1/\cos^2 t = 1 + \tan^2 t > 0$, donc strictement croissante dans l'intervalle ouvert $]-\pi/2, \pi/2[$ (fig. 8).

Le tableau de variation ci-dessus montre que, pour tout couple de nombres réels u, v tels que $u^2 + v^2 = 1$, il existe un nombre réel, et un seul, t dans l'intervalle $]-\pi, \pi]$ tel que :

$$\cos t = u, \quad \sin t = v.$$

Par suite, l'application $t \mapsto e^{it}$ est un homomorphisme surjectif de R sur U dont le noyau est le sous-groupe $2\pi\mathbf{Z}$ constitué des multiples entiers de 2π ; autrement dit,



Graphes des fonctions sinus et cosinus

t	0		$\pi/2$		π		$3\pi/2$		2π
cos	1	↘	0	↘	−1	↗	0	↗	1
sin	0	↗	1	↘	0	↘	−1	↗	0

deux nombres réels ont les mêmes cosinus et sinus si et seulement s'ils diffèrent d'un multiple entier de 2π.

Fonctions circulaires réciproques

Les fonctions circulaires n'étant pas monotones dans R tout entier, il ne sera possible de définir des fonctions réciproques que si l'on se restreint à des intervalles sur lesquels ces fonctions sont strictement monotones. On définit ainsi les fonctions Arc sinus, Arc cosinus et Arc tangente comme fonctions réciproques de la restriction du sinus à $[-\pi/2, \pi/2]$, de la restriction du cosinus à $[0, \pi]$ et enfin de la restriction de la tangente à $]-\pi/2, \pi/2[$ respectivement. Ainsi, Arc sin est une bijection strictement croissante de $[-1, +1]$ sur $[-\pi/2, \pi/2]$ et :

$$\begin{array}{ccc} y = \text{Arc}\sin x & \Leftrightarrow & x = \sin y \\ x \in [-1, +1] & & y \in [-\pi/2, +\pi/2]\,; \end{array}$$

Arc cos est une bijection strictement décroissante de $[-1, +1]$ sur $[0, \pi]$ et :

$$\begin{array}{ccc} y = \text{Arc}\cos x & \Leftrightarrow & x = \cos y \\ x \in [-1, +1] & & y \in [0, \pi]\,; \end{array}$$

Arc tan est une bijection strictement croissante de R sur $]-\pi/2, +\pi/2[$ (cf. fig. 8) et :

$$\begin{array}{ccc} y = \text{Arc}\tan x & \Leftrightarrow & x = \tan y \\ x \in \mathbf{R} & & y \in]-\pi/2, \pi/2[. \end{array}$$

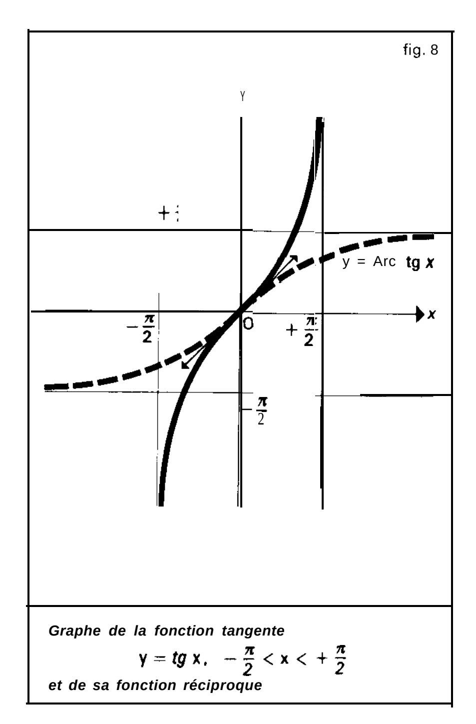
Le théorème des fonctions réciproques permet de calculer les dérivées de ces fonctions. On a :

$$(\text{Arc}\sin x)' = \frac{1}{\sqrt{1-x^2}}, \quad -1 < x < +1,$$

$$(\text{Arc}\cos x)' = -\frac{1}{\sqrt{1-x^2}}, \quad -1 < x < +1,$$

$$(\text{Arc}\tan x)' = \frac{1}{1+x^2}.$$

Ces fonctions sont d'usage constant en calcul intégral et permettent d'écrire le



Graphe de la fonction tangente

$y = \operatorname{tg} x, \quad -\frac{\pi}{2} < x < +\frac{\pi}{2}$

et de sa fonction réciproque

nombre π comme une intégrale définie, par exemple :

$$\frac{\pi}{2} = \int_0^1 \frac{dx}{\sqrt{1-x^2}}$$

Trigonométrie complexe

Les fonctions hyperboliques et les fonctions circulaires s'étendent au domaine complexe de manière naturelle, soit en utilisant des développements en série, soit (ce qui revient au même, puisque e^z est défini comme somme d'une série) au moyen des formules :

$$\operatorname{ch} z = \frac{e^z + e^{-z}}{2}, \qquad \operatorname{sh} z = \frac{e^z - e^{-z}}{2},$$

$$\cos z = \frac{e^{iz} + e^{-iz}}{2}, \qquad \sin z = \frac{e^{iz} - e^{-iz}}{2i},$$

qui mettent en évidence, par passage au domaine complexe, les liens étroits qui existent entre la trigonométrie hyperbolique et la trigonométrie circulaire. Ces fonctions sont analytiques dans tout le plan, et on a des relations du type :

$$\operatorname{ch} z = \cos iz, \quad \sin iz = i \operatorname{sh} z,$$

valables en particulier pour z réel, qui permettent de déduire la trigonométrie hyperbolique de la trigonométrie circulaire et vice versa.

Fonction argument principal

On a vu que $t \mapsto e^{it}$ est une bijection de $]-\pi, \pi]$ sur U, et ainsi $t \mapsto (\cos t, \sin t)$ est une représentation paramétrique du cercle trigonométrique : mais la bijection réciproque n'est pas continue au point -1. En revanche, l'application $\varphi : t \mapsto e^{it}$ est un homéomorphisme de $]-\pi, \pi[$ sur $U - \{-1\}$, la bijection réciproque ψ étant définie par :

$$\psi(u) = 2 \operatorname{Arc} \tan \left[\frac{1}{i}\left(\frac{u-1}{u+1}\right)\right], \quad u \in U - \{-1\},$$

le nombre $(1/i)((u-1)/(u+1))$ étant réel. En effet, pour tout nombre réel t de l'intervalle $]-\pi, \pi[$, on a :

$$\tan \frac{t}{2} = \frac{\sin (t/2)}{\cos (t/2)} = \frac{1}{i} \frac{e^{it} - 1}{e^{it} + 1}$$

Tout nombre complexe $z \neq 0$ peut s'écrire de manière unique sous la forme :

$$z = |z| e^{it}, \quad t \in]-\pi, \pi];$$

rappelons que t s'appelle l'*argument principal,* noté Arg z, du nombre complexe $z \neq 0$. D'après ce qui précède, l'application :

$$z \mapsto \operatorname{Arg} z$$

est continue sur C R , complémentaire dans C de l'ensemble des nombres réels négatifs, aussi appelé plan fendu ; on appelle cette applicationfonction *argument principal.*

En outre, si x_0 est un nombre réel strictement négatif. on a :

$$\lim_{\substack{z \to x_0 \\ \operatorname{Im} z > 0}} \operatorname{Arg} z = -\pi, \quad \lim_{\substack{z \to x_0 \\ \operatorname{Im} z < 0}} \operatorname{Arg} z = +\pi,$$

ce qui montre que la fonction argument principal ne se prolonge pas en une fonction continue sur $\mathbf{C} - \{0\}$.

Mentionnons enfin l'important théorème suivant :

Théorème de relèvement. Soit f une application de classe C^p, $p \geqslant 0$, d'un intervalle I de R et à valeurs dans U. Alors, il existe une application φ, de classe C^p, de 1 dans R telle que, pour tout t,

$$f(t) = e^{i\varphi(t)};$$

En outre, deux fonctions continues satisfaisant à cette relation diffèrent d'une constante de la forme $2k\pi$, $k \in \mathbf{Z}$.

logarithmes complexes

Les tentatives pour étendre les logarithmes aux nombres négatifs, puis aux nombres complexes, sont à l'origine d'une controverse célèbre qui a opposé, pendant près d'un demi-siècle, les plus grands esprits mathématiques du XVIIIe siècle. Jean Bernoulli admettait implicitement l'existence des logarithmes des nombres complexes, par analogie avec le cas réel, et il les introduisait tout naturellement à propos de l'intégration des fractions rationnelles, comme primitives d'éléments simples de la forme $1/(z - a)$, $a \in \mathbf{C}$. Bernoulli soutenait que :

$$\ln(-1) = 0,$$

car son double $\ln 1 = \ln(-1)^2$ est nul ; il en résultait que $\ln x = \ln(-x)$ pour tout réel positif, puis que $\ln(-1) = 0$ puisque $(-1)^2 = 1$. Mais ces résultats étaient en contradiction avec d'autres formules obtenues par J. Bernoulli lui-même.

Leibniz, pour sa part, soutenait que les logarithmes des nombres négatifs ne peuvent être réels. Une mémorable correspondance s'ensuivit entre les deux mathématiciens, de 1700 à 1716.

Dès 1728, L. Euler eut le pressentiment qu'il fallait abandonner l'unicité de la détermination si on voulait développer une théorie non contradictoire des logarithmes des nombres imaginaires. Dans un remarquable mémoire de 1749, il expose une théorie complète, en montrant que tout nombre non nul a une infinité de logarithmes possibles. Pourtant cela ne convainquit pas d'Alembert, qui continua la polémique.

Par analogie avec le cas réel, on est donc conduit à se demander si on peut définir le « logarithme » d'un nombre complexe $\zeta \neq 0$, c'est-à-dire à chercher un nombre complexe z tel que $e^z = \zeta$. Remarquons tout de suite que, s'il existe un tel nombre complexe z_0, alors la périodicité de la fonction exponentielle dans le domaine complexe, formule (32), entraîne que tout nombre complexe de la forme :

$$z_0 + 2k\pi i, \quad k \in \mathbf{Z},$$

est aussi un « logarithme » de ζ. Ainsi, si on veut avoir une théorie des logarithmes dans le domaine complexe, il sera indispensable d'associer à tout nombre une *infinité* de logarithmes.

Pour étudier la fonction exponentielle, nous nous limiterons d'abord, vu sa périodicité, à une « bande » du plan complexe dans laquelle la partie imaginaire de z varie dans un intervalle semi ouvert de longueur 2π ; nous prendrons ici la bande B formée des nombres complexes $z = x + iy$ tels que $-\pi < y \leqslant +\pi$.

La relation :

$$z = x + iy \in B, \quad e^z = \zeta,$$

équivaut aux relations :

$$\zeta \in \mathbf{C}^*, \quad x = \ln|\zeta| \text{ et } y = \operatorname{Arg}\zeta.$$

L'unique solution dans B de l'équation $e^z = \zeta$ s'appelle *logarithme* principal de ζ et se note $\ln \zeta$. Ainsi, par définition, pour $\zeta \in \mathbf{C}^*$

$$\ln\zeta = \ln|\zeta| + i \operatorname{Arg}\zeta.$$

Toutes les autres solutions de l'équation $e^z = \zeta$ sont alors de la forme :

$$\lg\zeta = \ln\zeta + 2k\pi i, \quad k \in \mathbf{Z};$$

la formule d'addition ne s'applique pas toujours à la détermination principale du logarithme mais peut s'énoncer ici : si $\lg \zeta$ et $\lg \zeta'$ sont des logarithmes de ζ et ζ' respectivement, alors $\lg \zeta + \lg \zeta'$ est un logarithme de $\zeta\zeta'$.

La fonction $\zeta \longmapsto \ln \zeta$ est continue sur le plan fendu $\mathbf{C} - \mathbf{R}^-$; on l'appelle *détermination principale du logarithme*. C'est une fonction *analytique* (cf. FONCTIONS ANALYTIQUES - Fonctions d'une variable complexe, chap. 4).

5. Développements eulériens des fonctions transcendantes élémentaires

Dans son *Introductio in analysis infinitorum* (1748), L. Euler définit l'exponentielle complexe par la formule :

$$e^z = \lim_{n\to\infty}\left(1 + \frac{z}{n}\right)^n;$$

par suite, il considère la fonction exponentielle et les fonctions trigonométriques qui s'en déduisent comme des « polynômes de degré infini ».

En particulier :

$$\sin z = \lim_{n\to\infty} \frac{1}{2i}\left[\left(1 + \frac{iz}{n}\right)^n - \left(1 - \frac{iz}{n}\right)^n\right];$$

écrivant le polynôme du second membre comme un produit de facteurs du second degré et faisant tendre n vers l'infini, il obtient la relation :

$$\sin z = z \prod_{n=1}^{\infty}\left(1 - \frac{z^2}{n^2\pi^2}\right),$$

qui est le développement de $\sin z$ en produit infini (produit eulérien).

Cette formule, valable pour tout $z \in \mathbf{C}$, met en évidence les zéros de la fonction sinus. tout comme la décomposition d'un polynôme comme produit de facteurs du premier degré (théorème de d'Alembert-Gauss, cf. nombres COMPLEXES, chap. 2).

Par des procédés analogues, il obtient les développements :

$$\operatorname{cotg} z = \frac{1}{z} + \sum_{n=1}^{\infty} \frac{2z}{z^2 - n^2\pi^2}$$

$$\frac{1}{\sin^2 z} = \sum_{-\infty}^{+\infty} \frac{1}{(z - n\pi)^2}$$

$$\frac{1}{\sin z} = \frac{1}{z} + \sum_{n=1}^{\infty} (-1)^n \frac{2z}{z^2 - n^2\pi^2},$$

valables pour $z \neq k\pi$, $k \in \mathbf{Z}$.

Cette fois, les fonctions de gauche dans les formules apparaissent comme des « fractions rationnelles de degré infini » ; au second membre figure alors la somme des parties principales, en chacun de leurs pôles, de ces fonctions (généralisation de la décomposition d'une fraction rationnelle en éléments simples).

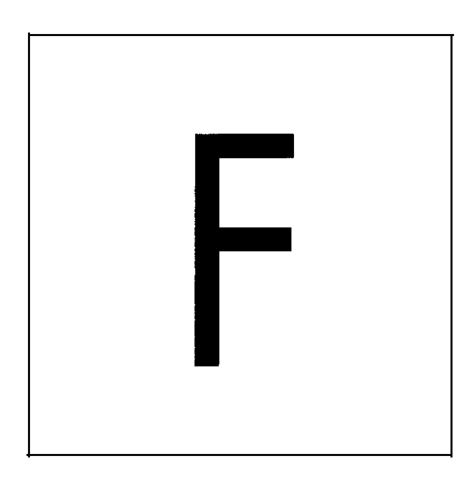
La théorie des fonctions analytiques fournit un cadre théorique permettant de

généraliser de telles formules (décompositions de Weierstrass et de Mittag-Leffler ; cf. FONCTIONS ANALYTIQUES - Fonctions d'une variable complexe, chap. 8).

JEAN-LUC VERLEY

Bibliographie

N. BOURBAKI, *Éléments de mathématique*, t. *IX : Fonctions d'une variable réelle,* Masson, Paris, 1982 / C. GILORMINI & G. HIRSCH, *Fonctions numériques d'une variable réelle,* Masson, 1980 / K. F. KLOPFENSTEIN, *Exponential and Logarithmic Functions*, Davies & Associates, Aurora (Colo.), 1991 / J. MARSDEN & A. WEINSTEIN, *Calculus I*, Springer-Verlag, 2ᵉ éd. 1985 / C. NAUX, *Histoire des logarithmes de Neper à Euler*, Blanchard, Paris, 1971.



FERMAT GRAND THÉORÈME DE

Il est rare qu'un théorème de mathématiques fasse la une des journaux internationaux. C'est pourtant ce qui est arrivé en juin 1993 quand un mathématicien britannique travaillant à l'université de Princeton, Andrew Wiles, a annoncé lors d'un colloque la preuve de la conjecture de Taniyama-Weil, formulée dans les années 1960 et dont on sait, depuis 1986, qu'une conséquence est le théorème de Fermat. C'est la connexion avec ce dernier, bien sûr, qui a fait naître l'émotion. Qu'a-t-il donc pour mériter une telle célébrité ? Un énoncé simple, une naissance quelque peu mystérieuse, une longue vie (près de trois cent cinquante ans) étroitement liée aux noms les plus célèbres des mathématiques occidentales. Et en apothéose, des avancées importantes qui soulignent les triomphes des mathématiques professionnelles du XXᵉ siècle.



L'énoncé

Un énoncé simple, donc, tout d'abord. Si on multiplie un nombre par lui-même, on obtient ce qu'on appelle un carré. Si on le multiplie encore une fois par lui-même, on obtient un cube ; encore une fois, une puissance quatrième, etc. Il existe des carrés qui sont la somme de deux autres carrés : par exemple $25 = 5 \times 5$ est la somme de $16 (= 4 \times 4)$ et de $9 (= 3 \times 3)$. Il y en a beaucoup d'autres (en fait une infinité), comme $4\,225 (= 65 \times 65)$ est égal à $1\,089 (= 33 \times 33) + 3\,136 (= 56 \times 56)$; à cause du fameux théorème de Pythagore, cela revient à dire qu'il existe des triangles rectangles avec des côtés entiers. Les choses se gâtent (ou deviennent plus intéressantes) dès qu'on passe des carrés aux cubes ou aux puissances supérieures. Il n'existe pas de cube somme de deux cubes, ni plus généralement de puissance d'exposant supérieur à 2, somme de deux puis-

sances de même exposant : autrement dit, l'équation $a^n + b^n = c^n$ n'a pas de solutions a, b, c en entiers non nuls dès que n est au moins égal à 3.

C'est cet énoncé d'apparence banale qu'un certain Pierre de Fermat, magistrat à Toulouse au XVIIe siècle, nota en marge d'un de ses livres de mathématiques. Il ajouta à l'énoncé, et la légende s'en est abondamment nourrie, que la marge était trop étroite pour contenir la merveilleuse démonstration qu'il en avait trouvée.

Pour les historiens modernes, il n'y a pas tant de quoi s'étonner. Fermat faisait partie de cette nouvelle classe intéressée par la science, et tout spécialement les mathématiques, qui prit un essor particulier au XVIIe siècle. Courtisans parfois, gens d'Église et plus souvent encore de robe, soldats occasionnels comme le fut Descartes, ils voyageaient, s'écrivaient beaucoup, s'informaient à travers toute l'Europe des derniers livres scientifiques disponibles, des instruments curieux, des observations astronomiques ou physiques et échangeaient des problèmes de toutes sortes. Fermat demanda à plusieurs de ses correspondants de prouver qu'un cube ne peut être somme de deux cubes ou une puissance quatrième somme de deux puissances quatrièmes. Il n'obtint jamais de réponse satisfaisante : même pour ces cas les plus simples, la preuve est difficile. Elle requiert un minimum de connaissances algébriques – qui restaient alors en Occident l'apanage d'un petit nombre, même parmi les amateurs de mathématiques – mais aussi une fine compréhension des problèmes spécifiques posés par les nombres entiers, que l'algèbre, s'appliquant à tous les nombres sans distinction, ne suffit pas à prendre en compte. Ce déploiement d'ingéniosité sur un tel énoncé parut à beaucoup des meilleurs mathématiciens de l'époque une perte de temps.

Quoi qu'il en soit, Fermat ne parla jamais du cas général de son « théorème » dans ses lettres ; une étude serrée des dates de ses recherches et de leur contenu indique qu'il comprit sans doute que sa démonstration n'était pas valide pour toutes les puissances. Fermat esquissa seulement dans une autre note une preuve pour l'exposant 4, et c'est par de maigres documents, des extraits de lettres, les fameuses notes, publiés par le fils du mathématicien après la mort de Fermat, que ses successeurs eurent accès à ses recherches. Au début du XIXe siècle, la plupart des énoncés de Fermat étaient soit munis de preuves, soit infirmés. À une exception près, celle qu'on sait. Il y avait déjà eu pourtant des tentatives, dont certaines dues à d'importants mathématiciens : Euler, Legendre, Dirichlet, Lamé avaient ainsi élucidé les cas des premiers exposants ($n = 3, 5, 7$) ; il suffit en effet de prouver le théorème pour 4 et pour les exposants premiers, c'est-à-dire non divisibles par un autre entier supérieur à 1, puisqu'il est alors vérifié automatiquement pour tous les multiples), chacun demandant un redoublement de persévérance et d'astuce.

Les premiers résultats généraux

Un changement important pour les mathématiques prit place dans le courant du XIXe siècle : dans les universités ou, en France, dans l'aire d'influence de l'École polytechnique apparurent des mathématiciens professionnels, chercheurs et enseignants, de plus en plus spécialisés. Les effets furent rapides : vers 1850, Ernst Eduard Kummer, professeur à l'université de Breslau (avant de devenir une des grandes personnalités de l'université de

Berlin) démontra que le théorème de Fermat est vrai pour tous les exposants premiers inférieurs à 100 (sauf trois possibles exceptions qui échappaient à sa méthode). Pour Kummer même, le théorème de Fermat n'était déjà plus qu'une curiosité ; son principal mérite était de montrer l'efficacité des nouveaux outils mis au point dans des perspectives bien plus vastes. Kummer et ses successeurs, Richard Julius Dedekind, Leopold Kronecker, notamment, cherchèrent à étendre les propriétés de l'arithmétique usuelle à d'autres familles de nombres et à généraliser les notions de divisibilité, de décomposition en facteurs premiers, et bien d'autres. À long terme, ce sont des notions fondamentales de l'algèbre structurale et de la théorie des nombres modernes qui étaient ici en gestation.

Kummer reçut pourtant un prix de l'Académie des sciences de Paris pour ses résultats sur l'équation de Fermat. Et ce n'est pas le moindre paradoxe de cette histoire : alors que les mathématiciens développent peu à peu des techniques complexes, posent et résolvent de nouveaux problèmes, plusieurs prix (à Paris. à Gottingen) sont proposés pour la résolution de cette énigme anodine déjà vieille de plus de deux siècles. Au cours du XX[e] siècle, un fossé grandissant va se creuser : des amateurs nombreux informés de l'existence du problème (et des prix associés) par les livres de popularisation, ignorant souvent tout des travaux mathématiques et historiques pertinents, partirent en quête de la preuve (élémentaire) qu'aurait pu trouver Fermat. Tandis que les mathématiciens travaillaient sur d'autres sujets, quitte à essayer au passage, lorsqu'elles pouvaient être pertinentes, leurs nouvelles approches : l'équation de Fermat faisait partie du folklore. Des raffinements théoriques du travail de Kummer et la puissance accrue des moyens de calcul sur ordinateur permettaient, il y a une dizaine d'années, de prouver le théorème de Fermat pour quelques centaines de milliers d'exposants premiers. Mais on avait aussi appris à connaître ses limites théoriques.

L'approche de Wiles

La dernière décennie a vu naître de nouveaux espoirs, le théorème de Fermat ayant été intégré dans plusieurs théories en plein développement. En améliorant des estimations analytiques très fines, Étienne Fouvry prouva en 1985 l'existence d'une infinité d'exposants premiers p pour lesquels l'équation de Fermat $a^p + b^p = c^p$ n'a pas de solutions, sauf peut-être si p divise le produit abc : c'était un des premiers résultats portant sur une famille infinie d'exposants premiers. D'autres approches bénéficièrent des avancées importantes des années 1960 en géométrie algébrique grâce aux travaux d'Alexandre Grothendieck, Jean-Pierre Serre, Pierre Deligne, John Tate et bien d'autres : cette géométrie réussit en particulier, au prix de grandes difficultés techniques, à garder trace du type de nombres (entiers, fractions, nombres réels, complexes. etc.) avec lesquels on travaille. Nous n'évoquerons ici que deux lignes de recherche, la seconde étant celle suivie par Andrew Wiles.

La première consiste simplement à interpréter l'équation $a^n + b^n = c^n$ comme celle d'une courbe plane $x^n + y^n = 1$ (en posant $x = a/c$, $y = b/c$ et en choisissant x et y comme coordonnées du plan). Le théorème de Fermat revient à montrer que cette courbe n'a aucun point dont les coordonnées soient entières ou fractionnaires (non nulles). En 1983, Gerd Faltings a démontré qu'une courbe de genre supé-

rieur ou égal à 2 n'a qu'un nombre fini de points à coordonnées fractionnaires (le genre est un invariant de la courbe lié au degré et aux singularités ; pour $x^n + y^n = 1$, le genre est $(n - 1)(n - 2) / 2$, donc supérieur ou égal à 2 dès que n est supérieur à 3). Le résultat de Faltings impliquait donc que le théorème de Fermat n'avait au plus qu'un nombre fini de solutions non nulles (à un facteur commun près) pour chaque exposant. La démonstration complète aurait été que ce nombre fini fût nul. Mais on n'en connaît pas de version effective, c'est-à-dire permettant de déterminer une borne pour les solutions éventuelles et, donc, de les rechercher systématiquement. Les essais pour rendre le résultat de Faltings effectif ont occupé beaucoup de spécialistes ces dix dernières années, et de nombreuses généralisations importantes de son théorème ont été découvertes au passage.

Une autre approche a conduit aux travaux d'Andrew Wiles. Elle consiste à relier l'étude de l'équation de Fermat, $a^n + b^n = c^n$, à celle de l'équation (*) $y^2 = x(x \quad a^n)(x + b^n)$. Celle-ci définit aussi une courbe plane (avec les coordonnées x et y) : mais, cette fois, une éventuelle solution à l'équation de Fermat déterminerait les coefficients de l'équation de la courbe et non un de ses points. Si a, b, c ne sont pas nuls, la courbe est de genre 1 et est un exemple de ce qu'on appelle une courbe elliptique. L'équation (*) en liaison avec le théorème de Fermat a été étudiée par Yves Hellegouarch dans les années 1970, mais c'est seulement au milieu des années 1980 qu'elle est revenue sur le devant de la scène lorsque Ken Ribet, motivé par une idée de Gerhard Frey et d'importants travaux de Serre, prouva que la courbe (*), si a, b, c ne sont pas nuls, contredirait une conjecture centrale des mathématiques, la conjecture de Taniyama-Weil. Autrement dit, si la conjecture de Taniyama-Weil était vraie, la courbe (*) ne pouvait exister, donc il ne pouvait y avoir de solutions non nulles à l'équation de Fermat. Cette relation contribua à renforcer la crédibilité du « théorème » de Fermat auprès des mathématiciens, car la conjecture de Taniyama-Weil, reliée à de nombreux autres phénomènes, semblait très convaincante. Comme l'a déclaré Andrew Wiles dans le journal de l'université de Princeton, « cela faisait du "théorème" de Fermat une conséquence d'un problème que les mathématiques ne pouvaient ignorer. Toute une architecture en dépendait ». C'est donc la preuve de cette conjecture (au moins la preuve d'une partie importante qui suffit pour l'application au théorème de Fermat) qu'a démontrée Wiles.

La conjecture de Taniyama-Weil, formulée dans les années 1960, prédit l'existence d'un dictionnaire précis entre les courbes elliptiques dont l'équation a des coefficients entiers ou fractionnaires et des fonctions périodiques spéciales dictionnaire proche de celui qui existe entre le cercle et les fonctions cosinus et sinus. Pour l'établir, Wiles met en œuvre des techniques variées des mathématiques actuelles (représentations galoisiennes. géométrie, cohomologie des courbes modulaires, systèmes d'Euler) et fait appel à des dizaines d'articles écrits par des chercheurs de nombreux pays dans les quinze dernières années.

Cette multiplicité des outils utilisés est en partie responsable de la durée et de la complexité des vérifications qui ont été nécessaires. À la fin de 1993, on ignorait si toutes les étapes de la preuve proposée pouvaient être confirmées, En octobre

1994, Wiles et Taylor produisaient une démonstration complète.

Quel est le rapport entre cette réalité du travail mathématique contemporain, d'une haute technicité. nécessitant l'intervention d'une communauté internationale dont seuls quelques noms ont pu être mentionnés ici, et l'effervescence qu'a suscitée le théorème de Fermat ? La réponse n'est pas simple. On pourrait se réjouir que l'énoncé de Fermat soit devenu un mythe et qu'il fournisse l'occasion de s'intéresser aux mathématiques d'aujourd'hui. Ou déplorer que les mathématiques soient tellement coupées de la culture ordinaire que seule la légende leur offre une voie d'accès. La réalité, y compris celle de la théorie des nombres contemporaine, liée de multiples façons non seulement à la géométrie mais aussi à la physique par exemple, n'est-elle pas plus riche et plus intéressante que la fable superficielle d'un génie isolé gardant jusque dans sa tombe le secret élémentaire d'un petit énoncé ? Les remarquables recherches de Fermat lui-même méritent d'être mieux appréciées. A fortiori celles qui ont suivi. Si mathématiques et culture s'accordaient, pour changer ?

CATHERINE GOLDSTEIN

Bibliographie

C. GOLDSTEIN, « Le Métier des nombres », in *Éléments d'histoire des sciences*, Bordas, Paris, 1989 ; « Conjectures en arithmétique », in *La Science au présent*, Encyclopædia Universalis, Paris, 1992 / J. ITARD, *Essais d'histoire des mathématiques*, Blanchard, Paris, 1984 / J. P. SERRE, *Cours d'arithmétique*, P.U.F.. Paris, 1977.

Cassette vidéo

Fermat's Last Theorem and its Proof (60 min). M.S.R.I., University of Berkeley, Berkeley, 1993.

FONCTION NOTION DE

Le terme de fonction a été introduit par Leibniz (1692) dans un contexte géométrique : il s'agit pour lui de portions de lignes droites qui dépendent d'un point variable sur une courbe, comme la tangente ou la normale. Jean Bernoulli, en 1698, a repris ce terme pour désigner une quantité X « composée d'une manière quelconque de x et de quantités données », où x désigne en l'occurrence l'ordonnée du point variable sur la courbe ; plus tard (1718), il proposera de noter fx une fonction d'une quantité variable x (pas forcément rattachée à un contexte géométrique). Le concept de fonction ainsi dégagé a servi de base à Euler pour son exposé de l'analyse mathématique (1748) d'un point de vue formel et non plus géométrique. Euler définit une variable comme une quantité qui admet toutes les valeurs possibles, par opposition à une constante, dont la valeur est fixée, et il représente géométriquement une variable par un axe ; ensuite, il définit une fonction d'une variable comme « une expression analytique composée d'une manière quelconque de cette quantité variable et de nombres ou de quantités constants ». Bien entendu, il ne précise pas ce qu'il entend par « expression analytique », ni la façon dont elle doit être « composée » : la suite du livre montre que c'est au moyen des opérations algébriques élémentaires éventuellement itérées indéfiniment (séries, produits infinis) et des opérations transcendantes comme log, exp, sin, cos ; suivant la manière dont elles sont composées, Euler classe les fonctions en algébriques et transcendantes, les fonctions algébriques étant elles-mêmes subdivisées en rationnelles et irrationnelles.

Les fonctions sont représentées graphiquement par des courbes dans le plan, où l'ordonnée est la valeur de la fonction lorsque l'abscisse est la valeur de la variable ; inversement, Euler se demande si une courbe donnée correspond à une fonction, et il est amené à distinguer les courbes « continues », graphes de fonctions définies par des expressions analytiques, et les courbes « discontinues » (ou « mixtes » ou « irrégulières »), réunion de morceaux qui correspondent à diverses fonctions. Par la suite, Euler sera amené à étendre le concept de fonction, en remarquant que les « fonctions arbitraires » intervenant dans la solution de l'équation des cordes vibrantes donnée par d'Alembert (1747) n'étaient pas nécessairement définies par des expressions analytiques, mais plutôt par un graphe obtenu par le « tracé libre de la main » ; c'est l'origine physique du problème qui conduisit Euler a cette définition très générale des fonctions. Daniel Bernoulli (1753) a donné une autre solution de l'équation des cordes vibrantes au moyen de séries trigonométriques, et il pensait que sa solution était aussi générale que celle de d'Alembert-Euler, ce qu'Euler a vivement contesté. On se trouvait donc en présence de deux notions de fonction : la conception formelle d'expression analytique, et la conception « ensembliste » plus générale de correspondance arbitraire ; le problème du rapport entre les deux notions se trouvait posé et sa solution devait attendre la fin du XIXe siècle.

Tandis que Lagrange *(Théorie des fonctions analytiques*, 1797), restait attaché à la définition des fonctions par des expressions analytiques (sa théorie est fondée sur l'utilisation des développements en séries entières), Fourier observa dès 1807 que la formule intégrale qui donne des coefficients du développement d'une fonction en série trigonométrique conservait un sens pour une fonction arbitraire du type de celles introduites par Euler (l'intégrale étant interprétée comme une aire dont l'existence est admise comme une évidence intuitive) ; il en déduisit, sans se poser de problème de convergence, qu'une fonction arbitraire pouvait toujours être représentée par une série trigonométrique. mais la notion de fonction arbitraire était encore extrêmement vague pour lui. Les travaux de Bolzano (18 17) et de Cauchy (1821) devaient apporter **un** peu de clarté sur cette notion, en introduisant la définition des fonctions continues (au sens moderne du terme, qui ne se réfère pas du **tout** aux « courbes continues » d'Euler) ; Cauchy développa une théorie de l'intégration des fonctions continues dégagée de l'intuition géométrique, et Dirichlet (1829) parvint à démontrer qu'une fonction continue monotone par par morceaux était effectivement représentée par sa série de Fourier. Les réflexions de Dirichlet pour l'extension à des fonctions plus générales le conduisirent à une conception des fonctions arbitraires beaucoup plus vaste que celle de ses prédécesseurs (pour qui il s'agissait essentiellement de fonctions continues par morceaux) : comme exemple de fonction discontinue, il donne la fonctionftelle que $f(x)$ vaille 1 pour x rationnel et 0 pour x irrationnel (1837) ; il pose alors le problème de l'intégration des fonctions arbitraires suffisamment générales (par exemple avec un ensemble rare de discontinuités) : ce devait être l'objet de travaux de Riemann (pour les fonctions dont l'ensemble de discontinuités est de mesure nulle) et de Lebesgue (pour une classe beaucoup plus large, stable par des passages à la limite simple).

P. Du Bois-Reymond montra en 1873 que, contrairement à la conviction des mathématiciens, la série de Fourier d'une

fonction continue ne converge pas nécessairement vers cette fonction. Mais Weierstrass parvint cependant à raccorder les deux notions de fonction (expression analytique et fonction arbitraire) en montrant qu'une fonction continue est toujours la somme d'une série de polynômes convergeant uniformément sur tout intervalle fermé et borné (1885). Le problème de la représentation analytique des fonctions discontinues allait être abordé en 1898 par Baire, qui caractérise les fonctions discontinues limites simples des fonctions continues (fonctions de classe 1), puis donne une classification des fonctions selon laquelle les fonctions de classe α sont des limites simples de fonctions des classes α' avec $\alpha' < \alpha$ (α et ci peuvent être des ordinaux transfinis) ; Lebesgue (1905) montra l'existence de fonctions de chacune des classes de Baire et aussi de fonctions échappant à la classification de Baire : de telles fonctions n'admettent pas de représentation analytique.

À côté des fonctions d'une variable réelle, que nous avons considérées jusqu'à présent, les mathématiciens ont aussi étudié les fonctions de plusieurs variables, dont le domaine de définition est une partie d'un espace $\mathbb{R}^n$. Par ailleurs, le calcul des variations conduisit à la notion de fonction dont la variable est une courbe (théorie des fonctions de lignes développée par V. Volterra) ; M. Fréchet (1904) et E. H. Moore (1905) étendirent ces conceptions en prenant la variable dans un ensemble arbitraire, et Fréchet (1909) eut même l'idée de considérer des fonctions non plus numériques, mais prenant leurs valeurs dans un ensemble quelconque : c'est la notion générale de fonction ou application utilisée de nos jours.

CHRISTIAN HOUZEL

FONCTION GAMMA → GAMMA FONCTION

FONCTION ZÊTA → ZÊTA FONCTION

FONCTIONS REPRÉSENTATION & APPROXIMATION DES

Il arrive très souvent que, dans les problèmes issus des mathématiques ou des autres sciences, les fonctions qui interviennent soient définies par des procédés qui ne permettent pas d'étudier de manière efficace leurs propriétés. C'est le cas des fonctions définies comme solutions d'équations fonctionnelles, d'équations différentielles ou intégrales, d'équations aux dérivées partielles, ou encore de problèmes variationnels. Ainsi, les fonctions exponentielles sont les solutions suffisamment régulières de l'équation fonctionnelle $f(x+y) = f(x)f(y)$, ou encore de l'équation différentielle $f'(x) = af(x)$.

On essaie alors de ***représenter*** une telle fonction f sous une forme plus efficace pour l'étude du problème posé (existence et unicité, variation de f, comportement asymptotique, dépendance de paramètres, approximation numérique, prolongement analytique...).

Quelques grandes méthodes se sont progressivement dégagées.

Il s'agit en premier lieu des *représentations continues,* obtenues à l'aide du calcul intégral. Dès le XVIIe siècle, le calcul des primitives a été utilisé pour représenter certaines fonctions. Ainsi, la fonction logarithme satisfait au problème de Cauchy $g'(t) = 1/t$, $g(1) = 0$, d'où :

$$\ln t = \operatorname{Log} t = \int_1^t \frac{du}{u}.$$

Cette relation permet de prouver l'existence du logarithme et, par suite, de l'exponentielle, ainsi que leur variation et leur comportement à l'infini (cf. EXPONENTIELLE ET LOGARITHME). Les représentations intégrales interviennent aussi sous la forme d'intégrales dépendant de paramètres, introduites par Leibniz. Les transformations de Fourier et de Laplace sont de ce type.

Il s'agit en second lieu des représentations *discrètes* et, au premier chef, des *développements* en *série entière.* Ainsi, la recherche des solutions sur R du problème de Cauchy $y' = y$, $y(0) = 1$, sous forme de série entière conduit à l'expression :

$$\exp x = 1 + \frac{x}{1!} + \dots + \frac{x^n}{n!} + \dots$$

Cette expression fournit l'existence de la fonction exponentielle et une approximation polynomiale par majoration du reste. Combinée avec l'équation fonctionnelle, elle permet aussi de calculer des valeurs approchées en un point; elle permet encore de prouver que la fonction exponentielle est prépondérante au voisinage de $+\infty$ sur les fonctions puissance ; enfin, elle permet le prolongement analytique au plan complexe par :

$$\exp z = 1 + \frac{z}{1!} + \dots + \frac{z^n}{n!} + \dots$$

On peut alors étudier les solutions sur $[0, +\infty[$ du problème de Cauchy $y' = ay$, $y(0) = A$, où A et a sont des nombres complexes, et établir la condition de stabilité en fonction du paramètre a, à savoir $\operatorname{Re} a < 0$.

Selon les problèmes, on est amené à utiliser d'autres types de séries : séries de Fourier pour les phénomènes périodiques, séries de polynômes orthogonaux...

On peut aussi utiliser des procédés *d'approximation par des suites de fonctions.* Ainsi, la méthode du pas à pas d'Euler et Cauchy (cf. équations DIFFÉRENTIELLES, chap. 7), appliquée à l'équation différentielle $y' = y$, $y(0) = 1$, conduit à la relation :

$$\exp t = \lim_{n \to +\infty} \left(1 + \frac{t}{n}\right)^n.$$

De même, la méthode des approximations successives de Cauchy-Picard (cf. équations DIFFÉRENTIELLES, chap. 1), appliquée à cette même équation, fournit à nouveau le développement en série de la fonction exponentielle.

Au XVIIIe siècle, les problèmes de représentation et d'approximation sont étudiés dans le cadre formel, le passage au domaine numérique étant traité de façon purement expérimentale, tandis qu'au XIXe siècle, *les problèmes de convergence* jouent un rôle central ainsi que la validité des *opérations sur les représentations* utilisées (opérations algébriques, dérivation, intégration, sommation...).

Les problèmes numériques conduisent à étudier non seulement la convergence mais aussi la *vitesse de convergence* et la stabilité. Ces nouvelles préoccupations conduisent, à la fin du XIXe siècle, à *diversifier les modes de convergence, à* diversifier les objets par lesquels on approche les fonctions et à rechercher des

procédés optimaux. L'ensemble des travaux sur ce sujet a joué un rôle moteur dans la constitution de l'analyse fonctionnelle au début du XX[e] siècle.

Enfin, dans les problèmes d'approximation, il arrive souvent qu'on veuille approcher non pas une fonctionfmais la valeur surf d'une ***forme linéaire*** donnée : intégrale de f, valeur de f ou d'une de ses dérivées en un point, coefficients de Fourier d'une fonction périodique f... Pour ce type de question, on se ramène bien entendu à approcher f par des fonctions plus simples, mais, ici, le point de vue est différent car l'étude de la rapidité de convergence et de l'optimisation sont spécifiques de la forme linéaire considérée.

L'enseignement de ces questions en France est conçu selon un plan rigide : étude des modes de convergence dans un cadre abstrait, validation des opérations sur les séries et les intégrales, représentation des fonctions et, enfin, résolution de problèmes. Cette démarche est contraire à la pratique scientifique où l'approfondissement théorique des modes de représentation et d'approximation va de pair avec l'étude des problèmes visés.

Les méthodes de représentation et d'approximation jouent un rôle central dans l'analyse mathématique. Elles sont présentées de façon synthétique dans les cinq premiers chapitres, qui renvoient pour plus de détails sur chacune des méthodes décrites aux divers articles d'analyse. Dans les trois derniers chapitres, nous approfondissons les problèmes d'approximation en abordant notamment les questions de stabilité et de vitesse de convergence, spécialement utiles en analyse numérique.



1. Convergences usuelles en analyse

Pour traiter des problèmes de représentation et d'approximation des fonctions, il est indispensable de préciser ce que l'on entend par ***l'écart*** de deux fonctions. Dans les cas les plus simples, on peut définir cet écart à l'aide d'une ***norme*** sur l'espace vectoriel E de fonctions considéré (cf. espaces vectoriels NORMÉS).

Normes usuelles

Considérons d'abord l'espace vectoriel $\mathcal{C}([a, b])$ des fonctions continues à valeurs complexes sur un intervalle 1 = [a, b].

Les trois normes usuelles sont :

- la norme de la ***convergence uniforme :***

$$N_\infty(f) = \|f\|_\infty = \sup_{t\in I} |f(t)|;$$

la norme de la ***convergence en moyenne :***

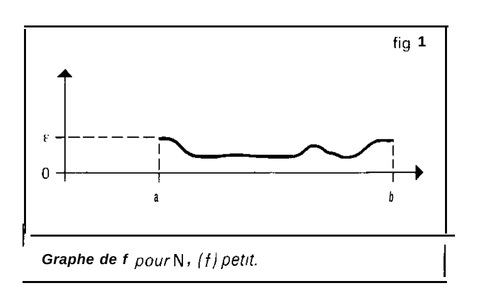
$$N_1(f) = \|f\|_1 = \int_a^b |f(t)|\,dt;$$

la norme de la ***convergence en moyenne yuadratique :***

$$N_2(f) = \|f\|_2 = \sqrt{\int_a^b |f(t)|^2\,dt}.$$

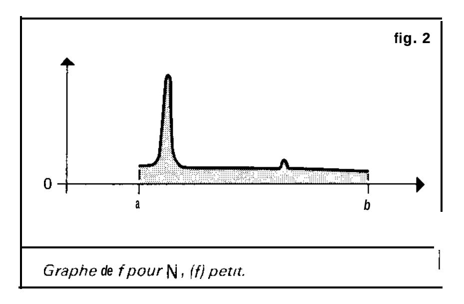
Soit f une fonction positive. Cette fonction f est petite au sens de la norme N_∞ si elle est petite partout ; cela signifie que son graphe est contenu dans une bande de hauteur petite (cf. fig. 1). La fonction f est petite au sens de la norme N_1 si l'aire de la partie hachurée sur la figure 2 est petite. Le cas de la norme N_2 s'y ramène en considérant la fonction f^2.

Il apparaît aussitôt que si f est petite pour N_∞, elle est aussi petite pour N_1 et N_2



Graphe de f pour $N_1(f)$ petit.

(si l'intervalle I n'est pas trop grand, cela va sans dire...), mais fpeut fort bien être petite pour N, sans l'être pour N_∞ (phénomène de « pointe », fig. 2) Mathéma-



Graphe de f pour $N_1(f)$ petit.

tiquement, ces considérations peuvent être précisées par les inégalités :

$$N_1(f) \leqslant (b-a)N_\infty(f),$$
$$N_2(f) \leqslant \sqrt{b-a}\,N_\infty(f),$$

qui expriment l'inégalité de la moyenne. De même, l'inégalité de Schwarz montre aussitôt que :

$$N_1(f) \leqslant \sqrt{b-a}\,N_2(f).$$

Ainsi, sur un intervalle compact, la convergence uniforme implique la convergence en moyenne quadratique, laquelle implique la convergence en moyenne. Mais les réciproques sont fausses, ce qui signifie que ces normes ne sont pas équivalentes.

Les trois exemples fondamentaux précédents se généralisent au cas où il est utile d'introduire un *poids*, c'est-à-dire une fonction π continue sur [a, b] à valeurs réelles strictement positives. On pose alors $N_\infty^\pi(f) = N_\infty(\pi f)$ et des définitions analogues pour les autres normes.

Ces exemples se généralisent aussitôt au cas des fonctions continues sur une partie compacte de $\mathbf{R}^n$ ou au cas des fonctions continues à support compact.

Dans la plupart des questions de convergence, il est indispensable que les espaces normés considérés soient complets (espaces *de Banach*). C'est le cas de l'espace E $=$ $\mathcal{C}([a,$ h]) muni de la norme N_∞. En revanche, ce même espace, muni de la norme N, ou de la norme N_2, n'est pas complet. La théorie de l'intégrale de Lebesgue permet de plonger respectivement E dans les espaces complets L'([a, b]) et $L^2([a,$ h]) des classes de fonctions intégrables ou de carré intégrable ; ces espaces sont les complétés de E pour les normes N, et N_2 (cf. espaces MÉTRIQUES, INTÉGRATION ET MESURE).

Les normes du type N_2 sont spécialement intéressantes pour deux raisons. La première est d'ordre mathématique : elles dérivent du produit hermitien :

$$(f,g) \mapsto \int_a^b \overline{f(t)}\,g(t)\,dt,$$

et on dispose donc de toutes les techniques hilbertiennes, très efficaces. L'autre tient au fait que ces normes se rencontrent dans de nombreux domaines de la physique mathématique : intégrales d'énergie, mécanique quantique, optimisation par la méthode des moindres carrés, processus stochastiques.

En revanche, l'étude directe des normes N_∞ est beaucoup plus délicate (cf. *infra*, chap. 7 et 8).

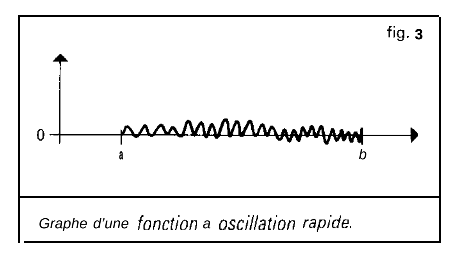
Normes et dérivation

Lorsqu'on s'intéresse à l'espace vectoriel $C^p([a,$ h]) **des** fonctions de classe C^p, il

convient souvent d'introduire des normes faisant intervenir les dérivés successives de f, par exemple :

$$f \mapsto N_\infty(f) + N_\infty(Df) + \dots + N_\infty(D^p f),$$
$$f \mapsto N_2(f) + N_2(Df) + \dots + N_2(D^p f).$$

Il est indispensable d'introduire les normes sur les dérivés successives, car la convergence uniforme de (f_n) vers 0 n'entraîne pas celle des dérivées : c'est le cas par exemple pour $f_n(t) = (1/n) \cos nt$ (il s'agit d'un phénomène d'oscillation rapide, fig. 3). En revanche, ce phénomène ne se



Graphe d'une fonction a oscillation rapide.

produit pas pour les fonctions de variable complexe (cf. théorème de Weierstrass, in FONCTIONS ANALYTIQUES Fonctions analytiques d'une variable complexe, chap. 5).

Ici encore, le cas des normes N_2 est plus simple à étudier. En outre, dans bon nombre de questions, on peut utiliser ces normes pour étudier les normes N_∞ grâce au résultat suivant : il existe des constantes α et β telles que :

$$N_\infty(f) \leqslant \alpha N_2(f) + \beta N_2(Df).$$

Les inégalités de ce type étendues à plusieurs variables jouent un rôle important dans la théorie des équations aux dérivées partielles ; elles ont été introduites par Sobolev (cf. *infra*, chap. 6).

Convergences définies par une famille de semi-normes

Le cadre des espaces normés ne suffit pas pour couvrir les besoins de l'analyse : certains types de convergence ne peuvent être décrits à l'aide d'une norme. Voici quatre exemples importants.

1. Convergence uniforme sur tout compact. Soit par exemple E l'espace vectoriel des fonctions continues sur R. La convergence uniforme sur tout compact d'une suite (f_n) vers 0 signifie que, pour tout entier p,

$$N_p(f_n) = \sup_{|t| \leqslant p} |f_n(t)| \to 0 \text{ pour } n \to +\infty.$$

2. Convergence uniforme de toutes les dérivées d'une fonction C^∞. Par exemple, si E est l'espace des fonctions de classe C^∞ sur $I = [a, b]$, on dit qu'une suite (f_n) tend vers 0 en ce sens si, pour tout entier p,

$$N_p(f_n) = N_\infty(D^p f_n) \to 0, \quad \text{pour } n \to +\infty.$$

3. Fonctions de classe C^∞ à décroissance rapide ainsi que toutes leurs dérivées. Ici on considère l'espace $\mathcal{S}$ des fonctions C^∞ sur R telles que, pour tout couple (s, k) d'entiers naturels :

$$(1 + t^2)^s \, |D^k f(t)| \to 0, \text{ pour } t \to +\infty.$$

On introduit la famille dénombrable de semi-normes :

$$N_{s,k}(f) = \sup_{t \in \mathbf{R}} (1 + t^2)^s |D^k f(t)|.$$

La convergence dans S d'une suite (f_n) vers 0 signifie que, pour tout couple (s, k), $N_{s,k}(f_n) \to 0$.

Cet espace joue un rôle fondamental dans la théorie de la transformation de Fourier, dans le cadre des distributions tempérées (cf. DISTRIBUTIONS, chap. 4).

Ces trois exemples se généralisent aux espaces $\mathbf{R}^n$.

Dans ce type de situation, où intervient une famille dénombrable de semi-normes, que l'on peut ranger en une suite

N_p, on peut décrire la convergence au moyen d'une *distance invariante par translation :*

$$d(g, h) = J(h - g),$$

où :

$$J(f) = \sum_{p=0}^{+\infty} \frac{1}{2^p} \inf(1, N_p(f)).$$

Les espaces métriques ainsi obtenus sont *complets* (cf. espaces MÉTRIQUES). Ce cas se distingue de celui des espaces normés par la perte de *l'homogénéité.*

4. *Convergence simple.* Soit par exemple E l'espace vectoriel des fonctions définies sur R à valeurs complexes. On dit qu'une suite (f_n) converge *simplement* vers 0 sur R si, pour tout point x de R, la suite *numérique* $(f_n(x))$ converge vers 0. On introduit les semi-normes (c'est-à-dire qu'un élément non nul peut avoir une semi-norme nulle) :

$$N_x(f) = |f(x)|.$$

Cette fois le cadre des espaces métriques ne suffit plus pour décrire ce type de convergence, car la famille (N_x) n'est plus dénombrable. D'ailleurs, contrairement aux apparences et à la terminologie (!), ce mode de convergence est assez pathologique car les propriétés stables sont peu nombreuses.

Les quatre exemples précédents relèvent de la théorie des *limites projectives* d'espaces semi-normés, qui se place dans le cadre général de la théorie des espaces vectoriels topologiques localement convexes (cf. espaces vectoriels TOPOLOGIQUES, chap. 1). Dans les trois premiers exemples, les familles d'espaces normés sont dénombrables et on obtient donc des *espaces de Fréchet*, *ce* qui n'est pas le cas de l'exemple 4.

Comparaison des convergences simple et uniforme

Il y a deux cas importants où la convergence simple implique en fait la convergence uniforme sur tout compact.

1. *Cas équilipschitzien* (théorème d'Ascoli classique). Soit A un espace métrique compact et (f_n) une suite d'applications de A dans C. On suppose :

a) la suite (f_n) converge simplement vers f sur A ;

b) la suite (f_n) est équilipschitzienne dans un rapport *k*, c'est-à-dire :

$$\forall n\ \forall x, y \in A,\ |f_n(x) - f_n(y)| \leqslant k\, d(x, y).$$

Alors, la fonction limite f est k-lipschitzienne et la convergence est uniforme sur A.

2. *Cas des suites monotone.~* (théorème de Dini). Soit encore A métrique compact et $f_n : A \to \mathbf{R}_+$. On suppose :

a) pour tout *n*, la fonction f_n est continue sur A ;

b) la suite (f_n) est décroissante, c'est-à-dire $f_{n+1}(x) \leqslant f_n(x)$ pour tout $x \in A$;

c) la suite (f_n) converge simplement vers 0.

Alors la convergence est uniforme sur A.

Convergences avec conditions sur les supports

Les convergences avec conditions sur les supports jouent un rôle important dans les problèmes liés au calcul intégral et à ses extensions (mesures de Radon et distributions).

Pour les *mesures*, considérons par exemple l'espace vectoriel E = K(R) des fonctions à valeurs complexes continues sur R et à support compact. On est amené à considérer les suites (f_n) d'éléments de E convergeant vers 0 au sens suivant :

a) f_n converge vers 0 uniformément sur **R** ;

b) il existe un intervalle compact K = [a, *b*] tel que, pour tout *n*, le support de f_n est contenu dans K.

On introduit donc les espaces vectoriels $\mathcal{K}_p(\mathbf{R})$ constitués des fonctions *f* dont le support est contenu dans $[-p, p]$, muni de la *norme :*

$$N_p(f) = \sup_{-p \leqslant t \leqslant p} |f(t)|.$$

La convergence d'une suite (f_n) vers 0 signifie alors que les f_n appartiennent à un *même* espace $\mathcal{K}_p$ et convergent vers 0 dans cet espace.

Pour la *théorie des distributions,* il convient de remplacer $\mathcal{K}(\mathbf{R})$ par l'espace 1)(R) des fonctions C^∞ sur R à support compact, et la condition (a) par la condition :

a') pour tout entier *k,* la suite $D^k f_n$ converge vers 0 uniformément sur R (cf. *supra,* exemple 2).

On introduit les espaces $\mathcal{D}_p(\mathbf{R})$ munis de la *distance* introduite ci-dessus dans l'exemple 2. La convergence d'une suite f_n vers 0 signifie alors que les f_n appartiennent à un *même* espace $\mathcal{D}_p$ et convergent vers 0 dans cet espace.

Les espaces vectoriels X(R) et $\mathcal{D}(\mathbf{R})$ sont *complets* (ce qui signifie ici que chacun des espaces SC,(R) ou $\mathcal{D}_p(\mathbf{R})$ est complet).

Ces deux exemples se généralisent facilement au cas des fonctions définies sur un ouvert U de $\mathbf{R}^n$, en remplaçant les intervalles $[-p$, p] par une suite de compacts (K,,) tels que K_p soit contenu dans l'intérieur de K_{p+1} et que U soit la réunion des K,, *(suite exhaustive de compacts de* U).

Ils relèvent de la théorie des *limites inductives* d'espaces semi-normés, qu'il n'est ni commode ni naturel de placer dans le cadre de la théorie des espaces vectoriels topologiques mais qu'il convient plutôt de décrire grâce au concept *d'espace vectoriel bornologique de type convexe* (cf. espaces vectoriels TOPOLOGIQUES, chap. 3).

2. Représentations par des intégrales

Forme intégrale des restes

Dans les problèmes de calcul différentiel, la forme intégrale des restes de développements en séries ou de développements asymptotiques est essentielle pour le contrôle de ces restes : formule de Taylor classique, formule interpolatoire de Lagrange-Hermite (cf. *infra*, chap. 4). À la formule de Taylor se rattache le *théorème de division des fonctions différentiables* : Soit *f* une fonction de classe C^p sur un intervalle 1 de centre *a* avec $f(a) = 0$; alors *f* peut s'écrire sous la forme :

$$f(x) = (x-a)g(x),$$

où $$g(x) = \int_0^1 uf'(a + (x-a)u)\,du.$$

Il en découle queg est de classe C^{p-1} et que, pour I compact, on a, pour tout entier $k \leqslant p - 1$, la majoration :

$$M_k(g) \leqslant \frac{1}{k+1} M_{k+1}(f),$$

où M,(h) désigne le maximum du module de la dérivée n-ième de *h* sur I.

Ce théorème se généralise aussitôt à plusieurs variables : Si $f(a) = 0$, avec $a = (a_1, \dots, a_n)$, alors, pour tout $x = (x_1, \dots, x_n)$, on a :

$$f(x) = (x_1 - a_1)g_1(x) + \dots + (x_n - a_n)g_n(x),$$

où les fonctions g, sont de classe C^{p-1} si *f* est de classe C^p. Ce théorème est à la base de la théorie des idéaux de fonctions différentiables (cf. CALCUL INFINITÉSIMAL Calcul à plusieurs variables, chap. 3).

Transformations de Fourier et de Laplace

Dans les problèmes d'analyse harmonique des phénomènes non périodiques (cf. analyse HARMONIQUE, chap. 3), on utilise la transformation de Fourier, réelle ou complexe. définie par la relation :

$$(1) \qquad \hat{f}(z) = \int_{-\infty}^{+\infty} f(t)\, e^{-2i\pi tz}\, dt$$

(ou des formes analogues selon les auteurs), et la formule d'inversion :

$$(2) \qquad f(t) = \int_{-\infty}^{+\infty} \hat{f}(x)\, e^{2i\pi tx}\, dx\,;$$

intuitivement, la formule (1) décompose le signal $t \mapsto f(t)$ suivant toutes ses composantes harmoniques (analyse harmonique du signal), tandis que la formule (2) permet de reconstituer ce signal f à partir de ses composantes (synthèse harmonique du signal).

Dans le cas des fonctions définies dans $[0, +\infty[$, qui interviennent dans l'étude des régimes non permanents. on utilise la transformation de Laplace :

$$\mathcal{L}f(z) = \int_0^{\infty} f(t)\, e^{-zt}\, dt$$

(cf. calcul SYMBOLIQUE). Cette transformation intervient aussi dans la résolution des équations différentielles à coefficients algébriques, notamment l'équation hypergéométrique (cf. calculs ASYMPTOTIQUES, chap. 6).

Les transformations de Fourier et de Laplace se généralisent aussi aux espaces $\mathbf{R}^n$ et jouent alors un rôle fondamental dans la théorie des équations aux dérivées partielles linéaires (cf. équations aux DÉRIVÉES PARTIELLES - Théorie linéaire) et plus généralement des équations de convolution. Elles jouent **aussi** un rôle de premier plan en calcul des probabilités, sous la forme des fonctions caractéristiques d'une loi de probabilité, dans la théorie des processus stochastiques du second ordre et en mécanique quantique.

On peut aussi rattacher à l'analyse harmonique la transformation de Mellin, définie par la relation :

$$\hat{f}(s) = \int_0^{\infty} f(t)\, t^{s-1}\, dt,$$

qui n'est autre que la transformation de Fourier sur le groupe multiplicatif $\mathbf{R}_+^*$ dont la mesure invariante est dt/t (cf. analyse HARMONIQUE). Cette transformation intervient notamment dans les problèmes multiplicatifs de la théorie des nombres (cf. théorie des NOMBRES - Théorie analytique) ; la fonction gamma d'Euler (cf. fonction GAMMA) joue ici **un** rôle central.

Emploi de la dualité

La dualité consiste à représenter une fonction comme une forme linéaire sur un espace E de fonctions convenablement choisi. Ainsi. toute fonction f de puissance p-ième intégrable définit une forme linéaire continue T_f sur l'espace L^q des fonctions de puissance q-ième intégrdble, pour $1/p + 1/q = 1$, par la formule :

$$T_f(\varphi) = \int_{\mathbf{R}} f(x)\, \varphi(x)\, dx\,;$$

lorsque $1 < p < +\infty$, on obtient ainsi toutes les formes linéaires continues sur L^q (cf. INTÉGRATION ET MESURE, chap. 4). Mais, en utilisant d'autres espaces fonctionnels, on obtient ainsi des objets plus généraux que les fonctions, Cela tient au fait que, dans de nombreux problèmes, les opérations de passage à la limite nécessitent de sortir du cadre de la théorie des fonctions. Considérons par exemple une

suite (φ_n) de fonctions continues positives sur R telles que, pour tout n,

$$\int_{\mathbf{R}} \varphi_n(t)\,dt = 1$$

(cf. DISTRIBUTIONS, fig. 1), et dont la masse se concentre à l'origine, c'est-à-dire que, pour tout $a > 0$,

$$\int_{|t| \geqslant a} \varphi_n(t)\,dt \to 0 \quad \text{pour } n \to +\infty.$$

Intuitivement, les fonctions φ_n tendent vers la « fonction » δ de Dirac, que les physiciens définissent par $\delta(0) = +\infty$, $\delta(x) = 0$ pour $x \neq 0$ et $\int \delta(t)\ dt = 1$. Mais il n'existe aucune fonction, au sens précis de ce concept en mathématiques, satisfaisant à ces relations. Il suffit de considérer, pour s'en convaincre, 26. On tourne alors la difficulté de la manière suivante : on démontre d'abord que, pour toute fonction f continue à support compact sur R,

$$(1) \qquad \int_{\mathbf{R}} f(t)\,\varphi_n(t)\,dt \to f(0).$$

À toute fonction φ continue sur R, on associe la forme linéaire :

$$\mathrm{T}_\varphi : f \mapsto \int f(t)\,\varphi(t)\,dt$$

sur l'espace vectoriel E = $\mathcal{K}(\mathbf{R})$ des fonctions continues à support compact. La relation (1) s'interprète alors de la manière suivante :

$$\mathrm{T}_{\varphi_n} \to \delta,\ n \to +\infty,$$

où δ est la forme linéaire sur E telle que f-f(O). En outre, l'application $\varphi \mapsto \mathrm{T}_\varphi$ permet de plonger l'espace $\mathcal{K}(\mathbf{R})$ dans son dual topologique, à savoir, l'espace $\mathcal{M}(\mathbf{R})$ des mesures de Radon sur R (cf. INTÉGRATION ET MESURE). Lorsque E est l'espace 'B(R), on obtient les distributions (cf. DISTRIBUTIONS).

Le principal intérêt de ces généralisations de la notion de fonction est de fournir un cadre théorique permettant d'opérer en toute sécurité sur les représentations (cf. *infra*, chap. 5). L'utilisation des transformations de Fourier et de Laplace pour le calcul symbolique et les équations aux dérivées partielles est à cet égard exemplaire.

Noyaux de convolution

Les noyaux de convolution constituent un autre exemple très intéressant de représentation intégrale. Ils interviennent d'abord dans la résolution des équations différentielles à coefficients constants avec second membre $\mathrm{P(D)}f = b, f(0) = 0$; si on introduit la *solution élémentaire* E définie par la relation P(D)E $= \delta$, où δ est la mesure de Dirac, alors $f = \mathrm{E} * b$, c'est-à-dire :

$$f(x) = \int_0^x \mathrm{E}(x-t)\,b(t)\,dt.$$

Cette méthode s'applique aussi aux équations aux dérivées partielles à coefficients constants (cf. équations aux DÉRIVÉES PARTIELLES - Théorie linéaire). En particulier, dans le cas du laplacien en dimension trois, on a :

$$\Delta \frac{1}{\|x\|} = -4\pi\delta\,;$$

la solution élémentaire est donc E(s) = $1/4\pi\|x\|$. Le potentiel V, solution de l'équation de Poisson AV $= - 4\pi\mu$, où μ est à support compact, est donc donné par l'intégrale de convolution :

$$\mathrm{V}(x) = \iiint_{\mathbf{R}^3} \frac{\mu(y)}{\|x-y\|}\,dy.$$

De même, la résolution du problème de Dirichlet pour le cercle fait intervenir le

noyau de Poisson (cf. POTENTIEL ET FONCTIONS HARMONIQUES, chap. 1).

L'importance des noyaux de convolution s'explique par le fait qu'ils peuvent décrire tous les problèmes linéaires invariants par translation de la variable. Plus précisément. soit f une fonction continue ; alors, l'application u_f qui, à tout élément φ de $\mathcal{D}$, associe $f * \varphi$ est une application linéaire continue de $\mathcal{D}$ dans l'espace $\mathcal{E}$ des fonctions indéfiniment dérivables. En outre, elle est invariante par translation sur la variable, c'est-à-dire qu'elle commute à tous les opérateurs de translation T_a définis par $T_a\varphi(x) = \varphi(x \quad a)$.

Le problème central est alors de savoir si, inversement, toute application linéaire continue invariante par translation est de la forme précédente. La réponse est négative dans le cadre de la théorie des fonctions (penser par exemple à l'application identique), mais elle est positive dans le cadre de la théorie des distributions : si u : $\mathcal{D} \rightarrow \mathcal{D}'$, linéaire continue, commute aux translations, alors il existe une *distribution* T et une seule telle que $u(\varphi) = T * \varphi$. En outre, u se prolonge en une application linéaire continue $\tilde{u}$ de $\mathcal{E}'$ dans $\mathcal{D}'$ et, dans ces conditions, T = U(6). Dans tous les théorèmes de théorie du signal et d'automatique, ce résultat joue un rôle fondamental ; T s'appelle la « fonction » de transfert ou réponse impulsionnelle (Cf. calcul SYMBOLIQUE).

intégrales de contour

La formule intégrale de Cauchy, fondamentale dans la théorie des fonctions de variable complexe (cf. FONCTIONS ANALYTIQUES Fonctions analytiques d'une variable complexe, chap. 5), s'interprète aussi dans ce cadre ; on utilise cette fois la solution élémentaire de l'opérateur $\partial/\partial\bar{z}$:

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right),$$

à savoir :

$$\frac{a\,i}{\partial \bar{z}\, z} \quad \pi\delta.$$

La méthode se généralise aux fonctions analytiques de plusieurs variables complexes.

La formule intégrale de Cauchy conduit à étudier plus généralement la transformation de Hilbert :

$$\hat{f}(z) = \int_\gamma \frac{f(t)}{z - t}\, dt,$$

où f est une fonction continue définie sur un arc y du plan complexe. Cette transformation intervient dans le problème des moments.

Elle conduit aussi à la recherche de représentations des fonctions holomorphes et méromorphes par des intégrales de contour. Le cas des fonctions hypergéométriques en est un exemple significatif (cf. calculs ASYMPTOTIQUES chap. 6).

Noyaux intégraux

La représentation des fonctions par des noyaux intégraux s'applique aussi à l'étude des phénomènes linéaires non nécessairement invariants par translation. C'est le cas pour la résolution des équations linéaires à coefficients non constants.

On introduit à cet effet la résolvante de l'équation sans second membre :

$$x'(t) - A(t)\,x(t) = 0,$$

où, pour tout t, A(t) est une matrice carrée d'ordre n à coefficients complexes et $x(t)$ un élément de $\mathbf{C}^n$: c'est la fonction $R(s, t)$,

à valeurs vectorielles, solution du problème de Cauchy :

$$\frac{dR}{dt}(s,t) = A(t)\,R(s,t),\ R(s,s) = I_n$$

Dans ces conditions, l'unique solution du problème de Cauchy :

$$x'(t) = A(t)\,x(t) + b(t),\quad x(s) = 0,$$

est donnée par :

$$x(t) = \int_s^t R(\sigma,t)\,b(\sigma)\,d\sigma$$

(cf. équations DIFFÉRENTIELLES). Placée dans le cadre de la théorie des distributions, cette méthode s'étend aux équations aux dérivées partielles linéaires grâce au concept de noyau élémentaire (cf. équations aux DÉRIVÉES PARTIELLES - Théorie linéaire).

Les noyaux intégraux interviennent aussi dans la théorie des fonctions de Green (cf. équations DIFFÉRENTIELLES, chap. 3; POTENTIEL ET FONCTIONS HARMONIQUES), dans la résolution des équations intégrales (cf. équations INTÉGRALES) et, plus généralement, dans la théorie des opérateurs. Par exemple, si K est une fonction de carré intégrable sur $\mathbf{R}^2$, alors, pour toute $f \in$ L*(R), la fonction g définie presque partout par la relation :

$$g(x) = \int_{\mathbf{R}} K(x,y)\,f(y)\,dy$$

appartient encore à L'(R) et l'application $f \mapsto g$ est un endomorphisme continu de l'espace de Hilbert L^2. Mais. inversement, tout endomorphisme **continu** de L^2 n'est pas nécessairement de cette forme : on obtient ainsi seulement les opérateurs de Hilbert-Schmidt (cf. théorie SPECTRALE, chap. 2).

Pour obtenir un énoncé satisfaisant, le cadre de la théorie des fonctions ne suffit pas ; il convient d'utiliser la théorie des distributions. Soit K une distribution sur $\mathbf{R}^2$; alors, pour toute fonction $\psi \in$ ED(R), l'application :

$$T_\psi : \varphi \mapsto \iint_{\mathbf{R}^2} K(x,y)\,\varphi(x)\,\psi(y)\,dx\,dy$$

est une distribution sur R et l'application u_K qui à ψ associe T_ψ est une application linéaire de $\mathcal{D}(\mathbf{R})$ dans 'a'(R), continue en ce sens que, si φ_n tend vers 0 dans $\mathcal{D}$, alors T_{φ_n} tend vers 0 dans $\mathcal{D}'$. Le célèbre *théorème des noyaux* de Laurent Schwartz affirme que, réciproquement, toute application linéaire continue de $\mathcal{D}$ dans $\mathcal{D}'$ est de cette forme, c'est-à-dire peut être définie par un noyau distribution.

3. Représentations par des séries

Séries entières

La somme d'une série entière de rayon de convergence R est une fonction indéfiniment différentiable dans son disque de convergence, et les dérivées successives à l'origine sont données par la formule de Taylor.

Inversement, dans de nombreux problèmes, il est utile de représenter une fonction f de classe C^∞ par sa série de Taylor. Mais ici la situation est très différente selon qu'on se place sur le corps des nombres complexes ou sur celui des nombres réels. Dans le premier cas, la série de Taylor converge toujours vers la fonction dans tout disque où f est de classe C^∞ (d'ailleurs la dérivabilité suffit ; cf. FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe, chap. 2). Dans le cas du corps des réels, il existe des fonctions C^∞ dans un intervalle]— a, a[dont la série de Taylor converge, mais vers une autre fonction : c'est le cas de la fonction f définie par f(0) = 0, f(x) = exp(- 1/

x^2), exemple introduit par Cauchy dans son *Cours d'analyse* (1821). Il existe aussi des fonctions C^∞ dont la série de Taylor a un rayon de convergence nul : c'est le cas de la fonction :

$$x \mapsto \sum_{n=0}^{\infty} \frac{1}{n!} \exp(in^2 x).$$

Plus précisément, Émile Bore1 a montré que, pour toute suite (a,) de nombres complexes, il existe une fonction C^∞ sur R telle que, pour tout $n, f^{(n)}(0) = a_n$. Autrement dit, l'application de Taylor T de C"(R) dans l'anneau C[[X]] des séries formelles à coefficients complexes, définie par :

$$T : f \mapsto \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} X^n,$$

est *surjective.*

Il n'est donc pas possible, comme l'a tenté Lagrange dans la *Théorie des fonctions analytiques* (1797), de fonder le calcul différentiel sur le développement en série de Taylor.

Ce phénomène est à l'origine du concept de fonction analytique réelle : ce sont les fonctions de classe C^∞ dans un intervalle ouvert I de R développables en série de Taylor au voisinage de chaque point a de I. Ces fonctions peuvent être caractérisées parmi les fonctions C^∞ dans 1 à l'aide d'inégalités du type suivant, portant sur la rapidité de croissance des modules des dérivées successives : pour **tout** intervalle compact K contenu dans 1, il existe des constantes C_K et r_K telles que l'on ait, pour tout entier n :

(1) $$M_{K,n}(f) \leqslant C_K\, n!\,(r_K)^n,$$

où : $$M_{K,n}(f) = \sup_{t \in K} |D^n f(t)|;$$

ce résultat est dû à Pringsheim.

Bien entendu, en tout point a $\in$ I, l'application de Taylor T_a est *injective* sur le sous-espace vectoriel $C^\omega(I)$ des fonctions analytiques dans I. Plus généralement, on dit qu'un sous-espace vectoriel V de C"(I) est *quasi analytique* si la restriction de T,, à V est injective.

La condition (1) conduit plus généralement à considérer une suite croissante $b = (\beta_n)$ tendant vers + ∞ et logarithmiquement convexe, c'est-à-dire vérifiant $\beta_n^2 \leqslant \beta_{n-1}\,\beta_{n+1}$, et à introduire la sous-algèbre V_b de C"(I) constituée des fonctions f telles que l'on ait :

(2) $$M_{K,n}(f) \leqslant C_K\, \beta_n (r_K)^n.$$

L e *théorème d e Denjoy-Carleman* affirme que V_b est quasi analytique si et seulement si la série de terme général $(\beta_n)^{1/n}$ est divergente.

Si V_b est quasi analytique, ses éléments ont des propriétés très rigides : en particulier, le principe du prolongement analytique est valable et $V_b \cap \mathcal{D}(I)$ est réduit à $\{0\}$. Au contraire, si V_b n'est pas quasi analytique, $\mathcal{D}_b(I) = V_b \cap$ 'D(I) est dense dans $\mathcal{D}(I)$ et les propriétés de V_b ressemblent à celles de D(I). L'exemple le plus intéressant est fourni par les classes de Gevrey, où $\beta_n = (n!)^s$, $s > 1$, auquel cas V_b se note 'D,(I) ; cet espace étant muni d'un type de convergence analogue à celui de 'D(I), son dual topologique est constitué des ultradistributions : il est analogue à $\mathcal{D}'(I)$ mais beaucoup plus grand. Ces espaces interviennent dans l'étude des problèmes aux limites des équations aux dérivées partielles, où l'on introduit aussi l'espace vectoriel des fonctionnelles analytiques, c'est-à-dire le dual topologique de $C^\omega(I)$, dont les propriétés diffèrent profondément de celles des distributions.

Séries de polynômes

Puisque les séries entières ne suffisent pas pour représenter des fonctions suffisamment générales (continues, de classe C^k), il convient de considérer des séries de polynômes. D'après le *théorème de Weierstrass* (1886), toute fonction continue sur un intervalle compact est développable en série uniformément convergente de polynômes. Le point de vue de Newton et de Lagrange se trouve ainsi justifié, mais dans un cadre théorique plus approfondi. Weierstrass considère d'ailleurs que ce résultat constitue une définition constructive de la notion de fonction continue.

Les séries de polynômes et de fractions rationnelles sont aussi utiles dans la représentation des fonctions d'une variable complexe lorsque le domaine de définition n'est plus un disque ou dans le cas où la fonction présente des singularités (cf. Développements eulériens, in EXPONENTIELLE ET LOGARITHME; Théorème de Runge, in FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe, chap. 8 et 9). Dans d'autres questions, il est commode d'utiliser des produits infinis (cf. Théorème de Runge, in FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe, chap. 9), ou des séries de Dirichlet.

Analyse algébrique

Les manipulations sur les séries entières et ces autres objets constituent le cœur de l'*analyse algébrique*, développée dans un cadre formel par Euler dans l'*Introductio in analysin infinitorum* (1748) et reprise dans le cadre du concept de convergence par Cauchy. L'analyse algébrique intervient dans la théorie des fonctions transcendantes élémentaires (cf. EXPONENTIELLE ET LOGARITHME) et, plus généralement, des fonctions spéciales (cf. fonctions de BESSEL; FONCTIONS ANALYTIQUES - Fonctions elliptiques et modulaire ; fonction GAMMA) ; elle intervient aussi dans les problèmes arithmétiques et combinatoires grâce au concept de série génératrice : séries entières pour les problèmes additifs, séries de Dirichlet pour les problèmes multiplicatifs (cf. analyse COMBINATOIRE, théorie des NOMBRES - Théorie analytique des nombres).

Séries de Fourier

Dans les problèmes d'analyse harmonique des phénomènes périodiques (de période 1 pour fixer les idées), on utilise le développement des fonctions périodiques en série de Fourier :

$$f(x) = \sum_{n \in \mathbb{Z}} \alpha_n(f)\, e^{2i\pi nx},$$

où le coefficient de Fourier α_n est donné par :

$$\alpha_n(f) = \int_0^1 f(t)\, e^{-2i\pi nt}\, dt.$$

La problématique est ici la même que pour la transformation de Fourier. Les deux cadres théoriques principaux de validité de cette représentation sont l'espace des fonctions de carré intégrable sur [0, 1] et l'espace des distributions périodiques (cf. DISTRIBUTIONS ; SÉRIES TRIGONOMÉTRIQUES). C'est d'ailleurs l'étude des séries de Fourier qui a constitué une des principales motivations de la construction de l'intégrale de Lebesgue et de l'espace L^2.

La représentation par les séries de Fourier s'insère dans un cadre plus général : pour étudier un opérateur auto-adjoint d'un espace hilbertien de fonctions, on développe les fonctions en série de fonctions propres de cet opérateur, ce qui est possible lorsqu'elles constituent une base

hilbertienne ; le cas des séries de Fourier est relatif à l'opérateur de dérivation. Le cas des équations différentielles conduit à la théorie de Sturm-Liouville et à l'introduction des systèmes classiques de fonctions orthogonales (cf. équations DIFFÉRENTIELLES, chap. 3;polynômes ORTHOGONAUX). L'étude des équations intégrales, en particulier des équations de Fredholm, qu'il convient de placer dans le cadre des opérateurs compacts, utilise aussi ce type de représentation (cf. équations INTÉGRALES ; théorie SPECTRALE).

Cette même méthode s'applique encore à l'étude de problèmes relatifs aux équations aux dérivées partielles ; une idée centrale pour déterminer les fonctions propres consiste ici à se ramener au cas des équations différentielles par la méthode de séparation des variables.

4. Approximation par des suites

Le problème de la représentation des fonctions comme limites de fonctions plus simples est intimement lié à celui de l'approximation des fonctions, qui ne relève pas uniquement de problèmes d'analyse numérique mais constitue un mode de représentation utile dans des questions d'ordre théorique : problèmes d'existence et d'unicité, démonstration de théorèmes par passage à la limite (argument de densité...). Les procédés d'approximations sont très divers ; nous avons retenu cinq méthodes importantes.

Méthodes convolutives

On utilise l'effet *régularisant* de la convolution : si fest une fonction peu régulière et si φ est très régulière, alors $f * \varphi$ est aussi régulière que φ. En introduisant une *approximation de l'unité*, c'est-à-dire une suite (φ_n) de fonctions très régulières convergeant vers la mesure de Dirac δ (cf. *supra*, chap. 2), on approche f par des fonctions très régulières $f * \varphi_n = f_n$.

Le fait que les fonctions φ_n soient a valeurs positives joue ici un rôle essentiel. Ce procédé d'approximation est particulièrement intéressant : en effet, lorsquefest de classe C^p, non seulement f_n converge vers f, mais, pour tout $k \leqslant p$, les dérivés $D^k f_n$ convergent vers $D^k f$. En prenant pour φ_n des fonctions C^∞ à support compact, on obtient la densité des fonctions C^∞ dans la plupart des espaces fonctionnels classiques et même des espaces de distributions ; ainsi, pour tout ouvert U de $\mathbf{R}^n$, l'espace vectoriel $\mathcal{D}(U)$ des fonctions de classe C^∞ dans U à support compact est dense dans l'espace vectoriel K(U) des fonctions continues à support compact contenu dans U.

En prenant pour φ_n des fonctions polynomiales, on obtient une démonstration du théorème d'approximation polynomiale de Weierstrass ; on peut prendre par exemple le noyau de Landau :

$$\varphi_n(t) = \begin{cases} a_n(1-t^2)^n & \text{si } |t| \leqslant 1 \\ 0 & \text{si } |t| \geqslant 1, \end{cases}$$

où a_n est une constante de normalisation, c'est-à-dire telle que :

$$\int_{\mathbf{R}} \varphi_n(t)\,dt = 1.$$

La même méthode s'applique aussi à l'approximation uniforme des fonctions continues périodiques : on peut prendre par exemple les noyaux de Fejer (cf. SÉRIES TRIGONOMÉTRIQUES, chap. 1) ou de La Vallée Poussin :

$$\varphi_n(t) = a_n(\cos \pi t)^{2n},$$

où a_n est une constante de normalisation.

On notera que, en revanche, le noyau de Dirichlet (cf. SÉRIES TRIGONOMÉTRIQUES, chap. 1) ne convient pas pour toutes les fonctions continues, ce qui se traduit par le fait que la série de Fourier d'une fonction continue peut diverger ; cela tient au fait que ce noyau n'est pas positif. Mais la convergence a bien lieu si la fonction est suffisamment régulière, c'est-à-dire si le module de continuité (cf. *infra*, chap. 7) décroît assez vite.

Méthodes de troncature

Il s'agit d'approcher des fonctions définies sur un ouvert U de $\mathbf{R}^n$ par des fonctions à support compact contenu dans U. À cet effet, on utilise une suite exhaustive (K,) de compacts de U (cf. *supra*, chap. 1) et on construit une *troncature universelle,* c'estàdire une suite (χ_n) de fonctions telles que $0 \leqslant \chi_n \leqslant 1, \chi_n = 1$ sur K_n et $\chi_{n+1} = 0$ en dehors de K_{n+1}.

On approche alors f par la suite $f_n = f\chi_n$.

Dans des questions d'intégration, on peut prendre pour χ_n la fonction caractéristique de K,. Dans d'autres problèmes, il faut opérer moins brutalement. Par exemple, en prenant χ_n continue, on prouve que l'espace vectoriel X(U) des fonctions continues à support compact contenu dans U est dense dans l'espace vectoriel C,(U) des fonctions continues tendant vers 0 au bord de U (muni de la norme de la convergence uniforme) ou encore dans l'espace $L^2(U)$ des classes de fonctions de carré intégrable dans U (muni de la norme quadratique).

De même, on peut prendre χ_n de classe C^∞ ; on prouve alors que l'espace vectoriel $\mathcal{D}(\mathbf{R})$ des fonctions de classe C^∞ à support compact est dense dans l'espace vectoriel S(R) des fonctions de classe C^∞ à décroissance rapide.

On peut conjuguer les méthodes de convolution et de troncature pour montrer par exemple que l'espace vectoriel $\mathcal{D}(U)$ est dense dans l'espace vectoriel C(U) pour la convergence uniforme sur tout compact (cf. *supra,* chap. 1).

Méthodes itératives

II s'agit ici de méthodes où l'algorithme d'approximation de f est défini par une relation de récurrence. En voici trois exemples, dont les deux premiers sont d'importance capitale.

Méthode des approximations successives

Pour prouver l'existence et l'unicité et étudier les solutions d'équations portant sur des fonctions, on s'inspire du cas des équations numériques en généralisant la méthode des approximations successives au cadre abstrait des espaces métriques complets (cf. espaces MÉTRIQUES, chap. 3). Ce schéma s'applique en particulier au théorème d'inversion locale et des fonctions implicites (cf. CALCUL INFINITÉSIMAL - Calcul à plusieurs variables, chap. 2), aux équations différentielles grâce à la méthode de Picard (cf. équations DIFFÉRENTIELLES, chap. 1, 2 et 4), aux équations intégrales (cf. équations INTÉGRALES, chap. 2) et aux équations aux dérivées partielles.

Méthode de compacité

Pour étudier les solutions d'une équation du type f(x) = b, où f est une application continue d'un espace métrique compact K dans un espace de Banach E et où b est un élément donné de E, on construit une suite (x_n) de solutions approchées, c'est-à-dire telles que f (x,) tende vers b. Dans ces conditions, l'équation f(x) = b admet au moins une solution et, si l'on suppose en outre l'unicité d'une telle solution a, la suite (s,) tend vers a.

Cette méthode s'applique d'abord aux méthodes d'analyse numérique matricielle (méthode de Jacobi). Mais elle s'applique aussi aux problèmes d'analyse fonctionnelle. Pour démontrer qu'un ensemble K de fonctions définies sur un espace compact A à valeurs dans un espace E, de dimension finie pour simplifier l'énoncé, est compact, on utilise le théorème suivant, dû à Ascoli. On munit l'espace $\mathcal{C}(A, E)$ de la norme de la convergence uniforme. Il est alors équivalent de dire :

a) K est compact ;

b) K satisfait aux deux conditions suivantes :

K est fermé ;

- K est équicontinu, c'est-à-dire, $\forall s \in A$, $\forall \varepsilon > 0$, $\exists \eta$ tel que :

$$\forall x \in \mathbf{K}, \forall t \in \mathbf{A}, \quad (d(t,s) < \eta \Rightarrow \|x(t) - x(s)\| < \varepsilon).$$

Lorsque, par exemple, A est un intervalle $[a, b]$ de R et E = C, l'hypothèse d'équicontinuité est satisfaite lorsque K est constitué des fonctions x de classe C^1 telles que $\| Dx \|_\infty \leqslant \beta$, ou, plus généralement si Dx est de carré intégrable et si $\int_a^b | Dx |^2 \leqslant \beta$.

Méthode de Padé : fractions continues

On opère par analogie avec le développement des nombres en fraction continue (cf. approximations DIOPHANTIENNES, chap. 2).

Une telle méthode se décrit à l'aide du schéma général suivant.

Soit f une fonction définie sur un intervalle I de centre 0. Un développement de f en fraction continue est de la forme :

$$f(x) = a_0 + \cfrac{x}{a_1 + \cfrac{x}{a_2 + \cfrac{x}{a_3 + \dots}}},$$

qu'on écrit avec la notation :

$$f(x) = a_0 + \frac{x|}{|a_1} + \frac{x|}{|a_2} + \frac{x|}{|\dots} \frac{}{|\dots};$$

cela signifie que les fractions rationnelles :

$$r_n(x) = a_0 + \frac{x|}{|a_1} + \frac{x|}{|\dots} \frac{}{|a_n}$$

convergent vers f uniformément sur tout compact de I. Cette méthode a été étudiée systématiquement par Padé (1892). Déjà, Euler avait montré que :

$$e^x = 1 + \frac{x|}{|1} + \frac{x|}{|-2} + \frac{x|}{|-3} + \frac{x|}{|2} + \frac{x|}{|5} + \frac{x|}{|-2} + \frac{x|}{|-7} + x \dots$$

et avait déduit de développements analogues l'irrationalité de e. Lambert avait obtenu de même le développement en fraction continue de tanx et en avait déduit l'irrationalité de π. Avec le développement de l'analyse numérique, lié à celui des ordinateurs, ce type d'approximation des fonctions transcendantes élémentaires par des fractions continues a connu un regain d'intérêt. Ainsi, pour le calcul de e^x, avec $\ln\sqrt{2} \leqslant x \leqslant \ln\sqrt{2}$, la formule :

$$e^x = 1 + \cfrac{x}{\cfrac{x}{2} + \cfrac{k_0 + k_1 x^2 + k_2 x^4}{1 + k_3 x^2}}$$

avec k_0 = 1,000 000 000 000 327 1, k_1 = 0,107 135 066 4.56 464 2, k_2 = 0,000 594 589 869 018 8 et k_3 = 0,023 801 733 157 418 6, fournit une précision de $1{,}4 \cdot 10^{14}$. Pour obtenir la même précision avec des polynômes, il faudrait monter au degré 15.

Mais cette méthode s'applique aussi à des problèmes d'ordre plus théorique. Soit par exemple f une fonction méromorphe dans un disque de centre 0 et régulière à l'origine. Alors f est développable en série de Taylor :

$$f(z) = \sum_{n=0}^{\infty} a_n z^n ;$$

on peut espérer que, si l'on approchefpar des fractions rationnelles au voisinage de 0,

$$R_p(z) = \sum_{n=0}^{N_p} \lambda_n z^n \Big/ \sum_{n=0}^{M_p} \mu_n z^n,$$

alors les pôles de R_p vont approcher les pôles de f. Plus précisément, on calcule les coefficients A,, et μ_n en résolvant le système linéaire obtenu en identifiant les développements limités d'ordre $M_p + N$,, defet de R_p à l'origine et on écrit alors R_p sous forme de fraction continue. Cette méthode est souvent utilisée de manière expérimentale dans les applications et donne des résultats fiables.

Méthodes de projection

Les méthodes de projection peuvent être décrites par le schéma général suivant. On considère un espace E de fonctions, normé par exemple, et une suite croissante de sous-espaces vectoriels fermés E,, dont la réunion soit dense dans E. Pour chaque entier n, on se donne un projecteur continu u_n de E sur E,,. On essaie alors d'approcher un élément f de E par la suite $f_n = u_n(f)$.

Le cas des projecteurs orthogonaux est un exemple fondamental : l'espace vectoriel E est hilbertien, muni d'une base hilbertienne $(e_0, e_1$ ..., e,, . ..), E,, est le sous-espace engendré par (e,. . . . . e,,) et u_n est le projecteur orthogonal de E sur E,,. Dans ces conditions, f_n converge toujours vers f (cf. espace de HILBERT). Le cas des séries de Fourier est celui où E est l'espace vectoriel des fonctions l-périodiques de carré intégrable sur [0, 1] muni de la norme quadratique et où, pour tout n, e,,(t) = $\exp 2i\pi nt$ (cf. séries TRIGONOMÉTRIQUES). De même, lorsque E est l'espace vectoriel $L^2([-1, 1])$ et que e_n est le polynôme de Legendre L_n de degré n, on obtient ainsi un procédé canonique d'approximation en moyenne quadratique de f par des polynômes $u_n(f)$. Le cas plus général des développements en série de fonctions orthogonales entre dans ce cadre (cf. polynômes ORTHOGONAUX), ainsi que la méthode de Galerkine.

Ces résultats montrent tout l'intérêt des normes quadratiques N_2 sur les espaces de fonctions ; en effet, nous verrons au chapitre 8 que, pour ce type de problème, le comportement de la norme N_∞ est beaucoup moins agréable.

Emploi d'extremums

De nombreuses questions d'analyse fonctionnelle peuvent se ramener à la recherche d'extremums (problèmes variationnels, problèmes aux limites dans les équations aux dérivées partielles...) que l'on peut souvent schématiser de la façon suivante. Soit C une partie convexe d'un espace hilbertien E et x un élément de E. Il existe alors au plus un point z de C tel que $d(x, z) = d(x, C)$. Pour étudier l'existence de z et approcher ce point, on construit une suite (z_n) d'éléments de C telle que $\lim_{n \to \infty} d(x, z_n) = d(x, C)$. On montre que cette suite est une suite de Cauchy. La convergence de la suite (z_n) est alors assurée lorsque C est une partie fermée de E.

5. Interpolation et discrétisation

Position du problème

Ce sont les problèmes de tabulation numérique des fonctions transcendantes élémentaires (lignes trigonométriques, logarithmes) et, à partir du XVII[e] siècle, le calcul approché des intégrales et des dérivées qui ont été à l'origine du développement des méthodes interpolatoires. D'autre part,

dans de nombreux phénomènes continus intervenant en sciences physiques, décrits par exemple à l'aide d'une fonction, on ne connaît les valeurs de cette fonction qu'en un certain nombre de points qui correspondent aux mesures effectuées. Les problèmes issus de l'astronomie, en particulier la détermination de la trajectoire des planètes, ont aussi joué un rôle moteur, comme en témoignent notamment les travaux d'Euler, Lagrange, Legendre, Laplace et Gauss.

Dans tous les cas, le phénomène continu est remplacé par un phénomène discret. Plus précisément, supposons que le phénomène continu est décrit par une fonction numériquefd'une variable réelle. On se donne une subdivision S de l'intervalle $[\alpha, \beta]$, c'est-à-dire une suite croissante $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ de points de $[\alpha, \beta]$; le module de la subdivision S est le nombre :

$$A(S) = \sup_{0 \leqslant j \leqslant n-1} (\alpha_{j+1} - \alpha_j).$$

Lorsque $\alpha_0 = \alpha$, $\alpha_n = \beta$ et, pour tout j, $\alpha_{j+1} - \alpha_j = (\beta - \alpha)/n$, on dit que la subdivision est à pas constant, et $A(S) = (\beta \quad \alpha)/n$ s'appelle le pas de S.

Dans ces conditions, on associe à S la suite finie $(f(\alpha_0), \ldots, f(\alpha_n))$. Nous dirons qu'il s'agit d'une *discrétisation* de f. Le problème est alors de savoir dans quelle mesure on peut reconstituer f à partir des phénomènes discrétisés associés. À cet effet, on interpole la suite précédente par une fonction simple, par exemple une fonction polynomiale que nous noterons $P_S(f)$, de degré $\leqslant n$, prenant aux points α_j les mêmes valeurs que f. Dans le cas où S est à pas constant, cette méthode a été élaborée dès le XVIIe siècle, notamment par Newton et Gregory, dans le cadre du calcul des différences finies (cf. *infra*). L'étude du cas général a été ébauchée par Newton et reprise par Lagrange, Cauchy et Hermite.

Il s'agit alors d'estimer la différence $f - P_S(f)$, par exemple en majorant $N_\infty(f - P_S(f))$, et d'étudier l'influence de la taille de l'intervalle $[\alpha, \beta]$, du choix de S et de la régularité de f sur la qualité de l'approximation. 11 arrive souvent qu'on ne s'intéresse pas directement à f mais à la valeur d'une forme linéaire sur f. Pour ce type de problème, d'autres normes s'imposent : norme N_1 pour les intégrales, normes $f \mapsto N_\infty(f) + N_\infty(Df)$ pour les dérivées. Dans les phénomènes physiques, les normes N_2 interviennent souvent au titre de l'énergie.

On peut aussi se demander s'il est possible de reconstituerfcomme limite de fonctions $P_S(f)$ en raffinant les subdivisions, c'est-à-dire en faisant tendre le module A(S) vers 0. Mais cette question présente des difficultés (cf. *infra*, *Interpolation polynomiale*). C'est pourquoi on a recours à un procédé plus élaboré : on se donnefsur un intervalle $[a, b]$, on découpe cet intervalle en p intervalles de longueur $(b - a)/p$ et, sur chacun de ces intervalles notés $[\alpha, \beta]$, on approche f par $P_S(f)$. Le problème est alors de savoir comment on peut jouer sur les entiers n etp pour obtenir des approximations efficaces. Dans ce schéma plus élaboré, f est approchée sur $[a, b]$ par une fonction φ continue et polynomiale par morceaux. On peut imposer en outre à φ des conditions de régularité plus fortes aux $p - 1$ points de subdivision de $[a, b]$, ce qui conduit par exemple à la théorie des fonctions spline (cf. *infra*, *Interpolation polynomiale par morceaux)*. Les méthodes de discrétisation s'appliquent aussi à la recherche de solutions approchées d'équations différentielles, grâce à l'emploi d'équations aux différences finies (cf. équations DIFFÉREN-

TIELLES, chap. 7) ; plus récemment, le développement des calculs sur ordinateurs a permis de les appliquer avec succès aux équations aux dérivées partielles.

Nous commencerons par décrire les principales méthodes d'interpolation polynomiale et nous aborderons ensuite l'interpolation par les fonctions polynomiales par morceaux.

interpolation polynomiale

Interpolation de Lagrange

Étant donné un élément $b = (\beta_0, \dots, \beta_n) \in \mathbf{C}^{n+1}$, on veut étudier les polynômes P à coefficients complexes tels que, pour tout j,

$$\text{(1)} \qquad P(\alpha_j) = \beta_j.$$

À cet effet, on introduit l'application linéaire $u : \mathbf{C}[X] \to \mathbf{C}^{n+1}$ qui à tout polynôme P associe $(P(\alpha_0), \dots, P(\alpha_n))$. Les équations (1) s'écrivent alors sous la forme :

$$\text{(1')} \qquad u(P) = b.$$

Le noyau de u est constitué des polynômes P qui s'annulent en tout point α_j, c'est-à-dire qui sont divisibles par :

$$\text{(2)} \qquad N = \prod_{j=0}^{n} (X - \alpha_j).$$

Ce polynôme N. dont la donnée équivaut à celle du système S, joue un rôle fondamental dans la théorie de l'interpolation ; on l'appelle le *noyau interpolateur* associé à S. Le théorème de division euclidienne des polynômes montre que $\mathbf{C}[X] = \operatorname{Ker} u \oplus E_n$, où E_n est l'espace vectoriel des polynômes de degré $\leqslant n$.

L'application linéaire u induit un isomorphisme de E_n sur $\mathbf{C}^{n+1}$. Ainsi, il existe un polynôme P_b de degré $\leqslant n$ et un seul tel que, pour tout j, on ait $P_b(\alpha_j) = \beta_j$.

Les solutions de (1) sont alors les polynômes P de la forme $P = P_b + QN$, où $Q \in \mathbf{C}[X]$.

En particulier, pour tout i, il existe un polynôme L_i de E_n et un seul tel que :

$$\text{(3)} \qquad \begin{cases} L_i(\alpha_j) = 0 & \text{si } i \neq j \\ L_i(\alpha_i) = 1, \end{cases}$$

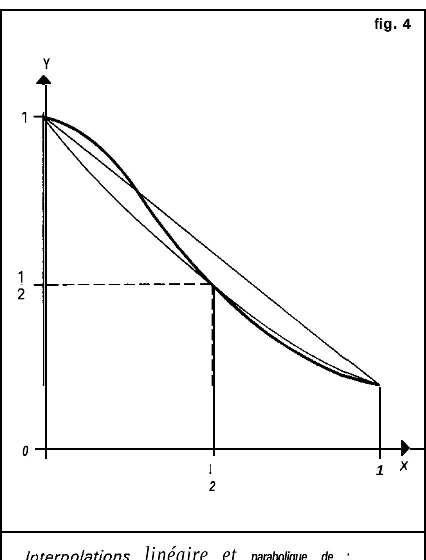
à savoir le polynôme :

$$\text{(4)} \qquad L_i = \prod_{j \neq i} \frac{X - \alpha_j}{\alpha_i - \alpha_j}$$

Dans ces conditions, on a :

$$\text{(5)} \qquad P_b = \sum_{i=0}^{n} \beta_i L_i.$$

Soit maintenant f une fonction continue à valeurs complexes sur $[\alpha, \beta]$, et $b = (f(\alpha_0), \dots, f(\alpha_n))$; le polynôme P_b s'appelle le *polynôme d'interpolation de Lagrange* de f associé à S et se note $L_S(f)$ (fig. 4). Ainsi. $L_S(f)$ est l'unique polynôme



Interpolations linéaire et parabolique de : $f(x) = \dfrac{1}{1+4x^2}$ *sur* [0,1].

de degré inférieur à n tel que, pour tout j,

(6) $$L_S(f)(\alpha_j) = f(\alpha_j),$$

et :

(7) $$L_S(f) = \sum_{i=0}^{n} f(\alpha_i) L_i.$$

En outre, l'application $u_S : f \mapsto L_S(f)$ est un projecteur de $\mathcal{C}([a, b])$ sur le sous-espace vectoriel E,,.

On peut estimer la précision de l'approximation de f par $L_S(f)$, grâce au théorème de division des fonctions différentiables (cf. *supra)* ; on démontre que, si fest de classe C^{n+1} sur [a, β], alors, pour tout $t \in$ [a, β], on a la majoration :

(8) $$|f(t) - L_S f(t)| \leqslant \frac{|N(t)|}{(n+1)!} M_{n+1}(f),$$

et, en particulier :

(9) $$\|f - L_S(f)\|_\infty \leqslant \frac{\|N\|_\infty}{(n+1)!} M_{n+1}(f).$$

Ces majorations ne peuvent pas être améliorées, car (8) et (9) sont des égalités lorsquefest le noyau interpolateur N.

Supposons maintenant que f est de classe C^∞ ; donnons-nous une suite (S_n) de subdivisions telle que A(S,) → 0 et notons plus simplement $L_n(f)$ le polynôme interpolateur de f associé à S_n. Dans ces conditions :

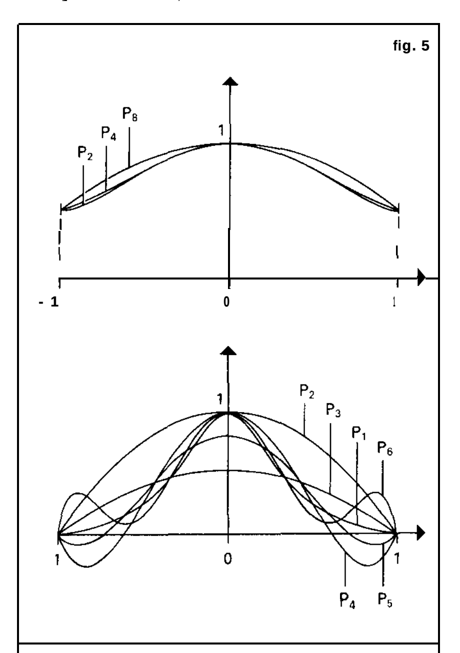
(9') $$\|f - L_n(f)\|_\infty \leqslant \frac{\|N_{n+1}\|_\infty}{(n+1)!} M_{n+1}(f),$$

où N_{n+1} est le noyau interpolateur associé à S,,, qui est de degré $n+1$.

Pour étudier la convergence de L,(f) vers f, on observe que :

$$\|N_{n+1}\|_\infty \leqslant (\beta - \alpha)^{n+1};$$

mais $M_{n+1}(f)$ peut croître très vite, si bien que la majoration (9') ne permet de prouver la convergence que sifest holomorphe dans un disque ouvert contenant l'intervalle [a, β]. D'ailleurs, dans le cas des subdivisions à pas constant $(\beta - \alpha)/n$, il peut arriver, contrairement à toute attente, que f soit analytique sur R mais que L,,(f) ne converge pas vers f sur $[\alpha, \beta]$. C'est le cas par exemple si $[a, \beta] = [-1, +1]$, f(t) $= 1/(1 + a^2t^2)$ (fig. 5) et a



Cette figure concerne l'interpolation des fonctions f: $x \to \frac{1}{1+x^2}$ et $g : x \to \frac{1}{1+8x^2}$ sur l'intervalle [−1, 1]. Les graphes de P_8 et de f sont pratiquement confondus. En revanche, pour la fonction g, les graphes de P_5 et de P_6 approchent très bien g au voisinage de 0, mais il apparait des oscillations importantes au voisinage des extrémités de l'intervalle (phénomène de Runge).

assez grand, par exemple a $\geqslant 2$. On observera qu'ici i/a et $-i/a$ sont des pôles de $f(z) = 1/(1 + a^2z^2)$, si bien que la condition d'holomorphie évoquée ci-dessus n'est pas réalisée. Ce type de phénomène a été découvert par Runge et Borel.

Pour remédier à ce défaut, on peut essayer de choisir les subdivisions S_n de telle sorte que $\| N_{n+1} \|_\infty$ soit minimale, ce qui a lieu lorsque N_n est le n-ième polynôme T_n de Tchebychev. Pour $[\alpha, \beta] = [-1, +1]$, on a :

$$(10) \qquad T_n(\cos\theta) = \frac{1}{2^n}\cos(n\theta),$$

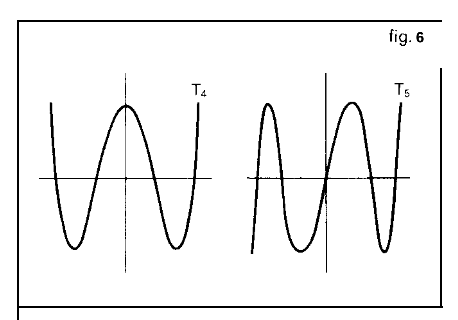
auquel cas :

$$\|T_n\|_\infty = \frac{1}{2^n},$$

et les points α_j sont :

$$\alpha_j = \cos\left[(2j+1)\frac{\pi}{2n}\right], \quad 0 \leqslant j \leqslant n-1$$

(fig. 6). Ainsi, les points α_j sont les



Polynômes de Tchebychev T_4 *et* T_5 *sur* $[-1, +1]$.

projections sur $[-1, +1]$ des sommets du polygone régulier à $2n$ côtés. Ces points s'accumulent donc aux deux extrémités de l'intervalle $[-1, +1]$. Nous verrons au chapitre 7, en utilisant une estimation beaucoup plus fine que (9′), que la convergence de $L_n(f)$ vers f est alors assurée sous des conditions très larges, par exemple si f est de classe C^1.

Cependant. cette convergence n'a pas lieu pour toutes les fonctions continues. En effet, d'après un théorème de Faber, quelle que soit la suite (S_n), il existe une fonction continue f telle que $\| L_n(f) \|_\infty$ tende vers $+\infty$. Ce résultat est un des exemples cités par Banach et Steinhaus comme étant à l'origine du théorème qui porte leur nom (cf. espaces vectoriels NORMÉS, chap. 4).

Interpolation de Hermite

Il s'agit du cas plus général où l'on impose au polynôme interpolateur des contacts d'ordre donné avec la fonction f aux points α_j. On se donne cette fois une suite $(\alpha_0, \ldots, \alpha_r)$ de points de $[\alpha, \beta]$, et une suite $(n_0, \ldots, n_r)$ d'entiers strictement positifs. On pose :

$$n = \sum_{j=0}^{r} n_j - 1$$

et :

$$N = \prod_{j=0}^{r} (X - \alpha_j)^{n_j}$$

Lorsque f est de classe C^{n+1} sur $[\alpha, \beta]$, il existe un polynôme $L_S(f)$ de degré $\leqslant n$ et un seul tel que, pour tout j, $0 \leqslant j \leqslant r$. et tout k, $0 \leqslant k \leqslant n_j - 1$, on ait :

$$(11) \qquad D^k L_s(f)(\alpha_j) = D^k f(\alpha_j)$$

En outre, pour tout $t \in [\alpha, \beta]$,

$$(12) \quad |f(t) - L_S(f)(t)| \leqslant \frac{|N(t)|}{(n+1)!} M_{n+1}(f).$$

Le polynôme $L_S(f)$ est le polynôme d'interpolation de Hermite associé à S. Le cas de Lagrange correspond à $n_j = 1$. tandis que la formule de Taylor correspond à $r = 1$ et $n_0 = n + 1$, c'est-à-dire :

$$N(t) = (t-a)^{n+1},$$

et :

$$L_S(f)(t) = T_a(f)$$

(cf. *supra*, chap. 3).

Dans le cas où $n_j = 2$, c'est-à-dire :

$$N = \prod_{j=1}^{r} (X - \alpha_j)^2,$$

on impose à $L_S(f)$ d'avoir un contact à l'ordre 1 en chaque point interpolateur α_j. Ce cas intervient notamment dans les méthodes d'optimisation pour le calcul approché des intégrales. La positivité du noyau N joue un rôle essentiel dans ce genre de question.

Différences divisées

L'expression du polynôme interpolatoire de Lagrange (formule (7)) a l'avantage de faire intervenir de manière symétrique les points de subdivision, mais elle est mal adaptée au calcul numérique, d'autant plus que l'adjonction de points interpolateurs supplémentaires oblige à recommencer entièrement les calculs. Newton avait procédé en utilisant une autre base de l'espace vectoriel des polynômes de degré $\leqslant n$, à savoir $D_0 = 1$, $D_1 = X - \alpha_0$, $D_2 = (X - \alpha_0)(X - \alpha_1)$, ..., $D_n = (X - \alpha_0)\ldots(X - \alpha_{n-1})$; ici la matrice de passage de la base canonique 1, X, ..., X^n à cette base est beaucoup plus simple que dans le cas de Lagrange car elle est triangulaire.

Pour obtenir la décomposition dans cette base du polynôme interpolateur $L_S(f)$, on introduit les *différences divisées* defrelatives à S, définies par les relations de récurrence :

$$\mathcal{F}_0(\alpha_0) = f(\alpha_0),$$

$$\mathcal{F}_1(\alpha_0, \alpha_1) = \frac{f(\alpha_0) - f(\alpha_1)}{\alpha_0 - \alpha_1},$$

$$\mathcal{F}_j(\alpha_0, \ldots, \alpha_j) = \frac{\mathcal{F}_{j-1}(\alpha_0, \ldots, \alpha_{j-1}) - \mathcal{F}_{j-1}(\alpha_1, \ldots, \alpha_j)}{\alpha_0 - \alpha_j}.$$

Dans ces conditions. on a la formule :

$$(13) \qquad L_S(f) = \sum_{i=0}^{n} \mathcal{F}_i(\alpha_0, \ldots, \alpha_i) D_i.$$

En outre, le reste est donné par la relation :

$$(14) \quad f(t) - L_S(f)(t) = N(t)\mathcal{F}_{n+1}(t, \alpha_0, \ldots, \alpha_n).$$

L'adjonction à S d'un point supplémentaire ne modifie pas les différences divisées précédentes. En outre, la différence divisée $\mathcal{F}_j$ est une fonction symétrique de α_0, ..., α_j, car :

$$\mathcal{F}_j(\alpha_0, \ldots, \alpha_j) = \sum_{k=0}^{j} \frac{f(\alpha_k)}{\prod_{h \neq k} (\alpha_h - \alpha_k)}.$$

Subdivisions à pas constant

Lorsque S est une subdivision de $[\alpha, \beta]$ à pas constant $h = (\beta - \alpha)/n$, il est commode d'étudier d'abord le cas où $h = 1$ et $\alpha = 0$, et donc $\beta = n - 1$; le noyau interpolateur s'écrit alors :

$$(15) \qquad N = X(X-1)\ldots(X-n+1).$$

Pour calculer le polynôme interpolateur $L_S(g)$, où g est définie sur $[0, n-1]$, on introduit l'opérateur de Bernoulli Δ, à savoir l'endomorphisme de C[X] défini par la relation :

$$(16) \qquad \Delta P(X) = P(X+1) - P(X).$$

L'étude de cet endomorphisme conduit à introduire une base de C[X] adaptée à Δ, définie par les relations de récurrence :

$$(17) \quad \begin{cases} N_0 = 1 \\ \Delta N_p = N_{p-1}, \quad N_p(0) = 0, \quad p \geqslant 1. \end{cases}$$

Ces polynômes sont donnés par les relations :

$$(18) \quad N_0 = 1, \; N_1 = X, \; N_2 = \frac{X(X-1)}{2!}, \ldots$$

$$N_p = \frac{X(X-1)\ldots(X-p+1)}{p!}.$$

Les polynômes N_p s'appellent *polynômes de Newton* ; on remarquera que

$N_p = (1/p\,!)\,D_p$. Dans ces conditions, tout polynome P se décompose sous la forme :

(19) $$P = \sum_{p=0}^{\infty} (\Delta^p P)(0) N_p \,;$$

c'est la formule de Newton-Gregory, analogue à la formule de Taylor qui est, elle, relative à l'opérateur de dérivation D, la base adaptée étant alors 1, X, ... $1/(p\,!)\,X^p$.

Appliquant la relation (19) à L,(g), on obtient :

(20) $$L_S(g) = \sum_{p=0}^{n} (\Delta^p g)(0) N_p .$$

Le calcul des différences successives dites non *divisées* $(\Delta^p g)(0)$ est très facile. Le cas d'une fonctionfdéfinie sur $[\alpha, \beta]$ se ramène au cas précédent par changement de variable affine, en introduisant g(u) $= f(\alpha + (\beta - \alpha)u)$.

On peut développer une théorie entièrement analogue en introduisant à côté de l'opérateur de Bernoulli *progressif* Δ, l'opérateur *régressif* ∇ défini par :

(21) $$\nabla P(X) = P(X) - P(X - 1)$$

et l'opérateur symétrique :

(22) $$\delta P(X) = P(X + \frac{1}{2}) - P(X - \frac{1}{2}).$$

Les opérateurs A, V et δ peuvent s'exprimer à l'aide des opérateurs de translation :

(23) $$T_h P(X) = P(X - h)$$

par les formules

(24) $$\Delta = T_{-1} - I,\ \ V=I-T,,\ \ 6 = T_{-\frac{1}{2}} - T_{\frac{1}{2}}\,;$$

on peut alors développer un calcul symbolique sur ces opérateurs et l'opérateur de dérivation D. Ainsi, la formule de Taylor traduit la formule symbolique $T_h = \exp hD$. Ce calcul, ébauché par Leibniz, a été systématisé par Lagrange et Laplace ; à la fin du XIX[e] siècle, il a été repris dans le cadre général de la théorie des opérateurs de convolution, c'est-à-dire des opérateurs qui commutent aux translations.

Généralisations

Le problème de l'interpolation se pose aussi pour les fonctions périodiques, auquel cas les fonctions polynomiales sont remplacées par des polynômes trigonométriques. Ces deux exemples se placent dans la théorie générale des *systèmes de Tchebychev* : on se donne **un** sous-espace E,, de dimension $n + 1$ de l'espace $\mathcal{C}([\alpha, \beta])$ qui est *régulier,* c'est-à-dire tel que tout élément de E,, qui s'annule en au moins $n + 1$ points est nul. Étant donné une base $\varphi_0, \varphi_1, \ldots, \varphi_n$ de E_n, on interpole f par une combinaison linéaire de ces fonctions. On peut montrer, en particulier, que l'espace vectoriel des solutions d'une équation différentielle linéaire sans second membre d'ordre $n + 1$ est régulier.

Interpolation polynomiale par morceaux

Théorie classique

On considère cette fois une fonction f continue sur un intervalle [a, b], que l'on découpe en p intervalles $[t_j, t_{j+1}]$ de longueur $(b - a)/p$ et, **sur** chacun de ces intervalles, on effectue une interpolation de Lagrange ou de Hermite de f, par des polynômes de degré $\leqslant$ n. En pratique, n est fixé et assez petit pour éviter les phénomènes du type de Runge. Il s'agit alors d'étudier la convergence de la suite

(φ_p) des fonctions polynomiales par morceaux ainsi obtenues vers la fonction f et la rapidité de convergence en fonction de la régularité de f.

Voici les deux exemples les plus importants.

1. ***Fonctions en escalier.*** Sur chaque intervalle $[t_j, t_{j+1}]$, on approche f par une constante, par exemple $f(t_j)$; ici $n = 0$. Grâce à la continuité uniforme de f sur $[a, b]$, on montre aussitôt que les fonctions en escalier ainsi obtenues convergent uniformément vers f sur ***[a, h].***

En outre, si f est lipschitzienne sur $[a, b]$, alors :

$$(25) \qquad N_\infty(f - \varphi_p) \leqslant \frac{b-a}{p}\lambda(f),$$

où $\lambda(f)$ désigne le rapport de Lipschitz de f, l'égalité étant atteinte si f est affine.

2. Fonctions affines par morceaux. Sur chaque intervalle $[t_j, t_{j+1}]$, on effectue une interpolation linéaire de f. On montre encore que les fonctions affines par morceaux φ_p ainsi obtenues convergent uniformément vers f sur [a, h]. En outre, si f est lipschitzienne de rapport k sur $[a, b]$, on a :

$$(26) \qquad N_\infty(f - \varphi_p) \leqslant \frac{k(b-a)}{2p},$$

l'égalité étant atteinte par exemple pour $[a, b] = [-1, 1]$ et $f(x) = |x|$.

Mais, cette fois, si f est de classe C^1, la rapidité de convergence est meilleure. Par exemple, si f' est lipschitzienne,

$$(27) \qquad N_\infty(f - \varphi_p) \leqslant \frac{(b-a)^2}{8p^2}\lambda(f'),$$

l'égalité étant atteinte par exemple si $[a, b] = [-1, 1]$ et $f(x) = x^2$.

Dans certains problèmes, comme le calcul approché des intégrales, on est amené à interpoler f par des fonctions polynomiales par morceaux de degré n plus élevé. Lorsque f est de classe C^{n+1}, on a alors :

$$(28) \quad N_\infty(f - \varphi_p) \leqslant \frac{(b-a)^{n+1}}{(n+1)!\,p^{n+1}} M_{n+1}(f).$$

On constate ainsi que la rapidité de convergence est gouvernée par la régularité de f.

Fonctions spline

Lorsque $n \geqslant 2$, le procédé précédent présente de nombreux inconvénients, car on approche des fonctions régulières f par des fonctions qui ne sont même pas de classe C^1. Il est donc intéressant d'approcher f par des fonctions φ satisfaisant aux deux conditions suivantes :

a) Sur chaque intervalle $[t_j, t_{j+1}]$, la fonction φ est polynomiale de degré $\leqslant n$;

b) Sur ***[a,*** b], la fonction φ est de classe C^{n-1} (conditions de recollement aux points t_j).

De telles fonctions sont appelées spline (tringle souple). L'espace vectoriel $S_n([a, h])$ de ces fonctions est de dimension $n + p$. Si on impose maintenant les $p + 1$ conditions interpolatoires relatives à la fonction f :

c) Pour tout j, $0 \leqslant j \leqslant p$, on a l'égalité $\varphi(t_j) = f(t_j)$, et il reste $n - 1$ paramètres encore libres.

Le cas $n = 2$ (fonctions paraboliques par morceaux) ne conduit pas à une théorie intéressante car il ne reste qu'un seul paramètre, ce qui ne permet pas d'imposer une condition de contact aux deux extrémités de l'intervalle ***[a, b].*** Au contraire, le cas $n = 3$ (fonctions spline cubiques) convient bien car il permet d'imposer la condition :

d) $\varphi'(a) = f'(a)$ et $\varphi'(b) = f'(h)$.

Les fonctions spline cubiques conviennent aussi pour l'interpolation des fonc-

tions périodiques. Dans ce cas, on impose à φ de se prolonger en une fonction périodique de classe C^2 sur R, ce qui conduit à remplacer la condition (d) par :

$d')$ $D_+\varphi(a) = D_-\varphi(b)$ et $D_+^2\varphi(a) = D_-^2\varphi(b)$, en désignant par D_+ et D_- les opérateurs de dérivation à droite et à gauche.

On démontre que les conditions (a), (b), (c), (d) ou (d'), déterminent φ de manière unique. Cette fonction se note $S_p(f)$ et s'appelle fonction spline cubique interpolant f à l'ordre p. Il faut cependant remarquer que le calcul explicite de φ nécessite la résolution d'un système d'équations linéaires assez compliqué. En revanche, la qualité de l'approximation est excellente :

si f est de classe C^1, on a :

$$N_\infty(f - S_p(f)) = o\left(\frac{1}{p}\right);$$

- si f est de classe C^2 :

$$N_\infty(f - S_p(f)) = O\left(\frac{1}{p^2}\right)$$

et, en outre :

$$N_\infty(f' - S_p'(f)) = o\left(\frac{1}{p}\right),$$

ce qui signifie qu'il y a aussi convergence uniforme pour la dérivée ;

si f est de classe C^3, alors, pour $k \leqslant 2$:

$$N_\infty(D^k f - D^k S_p(f)) = o\left(\frac{1}{p^{3-k}}\right);$$

enfin, si f est de classe C^4, alors, pour $k \leqslant 3$:

$$N_\infty(D^k f - D^k S_p(f)) = O\left(\frac{1}{p^{4-k}}\right).$$

Cependant, si f est de classe C^∞, ou même polynomiale de degré $\geqslant 4$, on ne peut pas améliorer la dernière estimation.

Un autre pôle d'intérêt des fonctions spline est relatif à leurs propriétés pour la norme quadratique. Plus précisément, si f est de classe C^2, pour tout élément φ de $S_{2,p}([a, b])$, on a :

$$\int_a^b |f'' - \varphi''|^2 \leqslant \int_a^b |f'' - S_p''(f)|^2 + \int_a^b |S_p''(f) - \varphi''|^2 ;$$

en particulier, parmi toutes les fonctions spline cubiques φ, c'est $S_p(f)$ qui approche le mieux f pour la norme de l'énergie. Cette propriété est à l'origine de la dénomination spline, car on peut montrer, grâce au principe de moindre action, que $S_p(f)$ réalise la position d'équilibre d'une tringle souple passant par les points interpolateurs et tangente à des droites données aux extrémités de l'intervalle. Ce procédé est depuis longtemps utilisé en dessin industriel. Les fonctions spline se sont révélées très utiles dans les problèmes de conception assistée par ordinateur.

La théorie des fonctions spline peut se généraliser aux spline d'ordre impair quelconque. Les propriétés d'approximation sont alors meilleures mais les calculs sont encore plus complexes que pour $n = 3$, si bien qu'elles sont peu utilisées en analyse numérique.

6. Opérations sur les représentations et les approximations

Nous avons déjà vu que l'emploi des représentations pour la résolution des problèmes nécessite de pouvoir opérer sur ces représentations : il s'agit non seulement des opérations algébriques (somme, produit...) mais aussi des opérations de passage à la limite (limites de suites, sommes de séries, intégration, dérivation). Ces

problèmes rentrent dans le schéma général d'interversions de passages à la limite. Nous commencerons par préciser les propriétés stables par passage à la limite uniforme pour les suites de fonctions, ce qui est étroitement lié au cas des séries, puis nous examinerons le cas des représentations intégrales. Nous terminerons par quelques indications sur les autres modes de convergence.

Suites de fonctions

Les trois problèmes les plus importants sont les suivants.

Problème 1. Continuité et passage à la limite. Soit (f_n) une suite de fonctions continues sur un espace métrique A à valeurs complexes, convergeant sur A vers une fonction f. La fonction f est-elle continue sur A?

La réponse est négative pour la convergence simple. comme le montre l'exemple A = [O. 1] et $f(x) = x^n$.

Théorème 1. Stabilité de la continuité. Si (f_n) converge uniformément versfsur tout compact de A, alors f est continue sur A.

Problème 2. Passage à lu limite dans les intégrales. On dispose de deux résultats très importants. Le premier est élémentaire.

Théorème 2. Passage à la limite dans les intégrales (cas compact).

Soit (f_n) une suite d'applications continues de [a, *b]* dans C qui converge uniformément vers f sur [a, h]. Alors :

$$\int_a^b \lim_n f_n = \lim_n \int_a^b f_n$$

(interversion des signes lim et $\int$).

Ce résultat s'étend si on remplace le segment $[a, b]$ par une partie compacte de $\mathbf{R}^m$.

Le second résultat a une portée beaucoup plus générale ; il se place dans le cadre de la théorie de l'intégrale de Lebesgue (Cf. INTÉGRATION ET MESURE).

Théorème 2 bis (théorème de convergence dominée). Soit (f_n) une suite de fonctions à valeurs complexes intégrables sur I et qui converge simplement vers f sur 1. On suppose qu'il existe une fonction $\varphi \geqslant 0$, intégrable sur 1, telle que l'on ait :

$$\forall n \quad |f_n| \leqslant \varphi.$$

Alors f est intégrable et on a :

$$\int_I \lim_n f_n = \lim_n \int_I f_n.$$

Ce résultat s'étend sans changement au cas où I est une partie localement compacte de $\mathbf{R}^m$.

Problème 3. Passage à la limite dans les dérivées. Soit (f_n) une suite de fonctions de classe C^1 sur un intervalle 1 de R, à valeurs complexes, convergeant vers f sur I. La fonction f est-elle de classe C^1 sur I et a-t-on $Df = \lim Df_n$?

L'exemple de I = [0, 1] et $f_n(x) = (\sin nx)/n$ montre que la convergence uniforme ne suffit pas. 11 faut donc faire des hypothèses sur la suite (Df_n).

Théorème 3. Passage à lu limite dans les dérivées. On fait les hypothèses suivantes sur la suite (f_n) :

a) f_n converge vers f uniformément sur tout compact de I ;

b) f'_n converge vers g uniformément sur tout compact de I.

Alors f est de classe C^1 sur I et $f' = g$, autrement dit :

$$D(\lim_n f_n) = \lim_n Df_n$$

Ce théorème s'étend aussitôt aux fonctions de classe C^p et C^∞ et au cas

des fonctions définies sur un ouvert U de $\mathbf{R}^m$.

On notera que, en revanche, dans le cas des fonctions analytiques d'une variable complexe, la convergence uniforme sur tout compact de f_n versf entraîne celle de Df_n vers Df (théorème de Weierstrass, cf. FONCTIONS ANALYTIQUES -Fonctions analytiques d'une variable complexe, chap. 5).

Séries de fonctions

On se donne une série Cu,, de fonctions définies sur un espace métrique A et à valeurs complexes.

On dit que cette série converge simplement (resp. uniformément) vers une fonction f si la suite des sommes partielles

$$S_p = \sum_{n=0}^{p} u_n$$

converge simplement (resp. uniformément) vers f.

Pour vérifier que la série converge uniformément, on utilise souvent la méthode des séries numériques majorantes. On dit qu'une série *numérique* β_n, $\beta_n \geqslant 0$, **est** une *série mujorante* sur A de la série de *fonctions* Σu_n si, pour tout n et pour $t \in A$, on a:

$$|u_n(t)| \leqslant \beta_n.$$

Si la série majorante converge, la convergence de la série Σu_n est uniforme sur A. Dans ces conditions :

$$\sum_{n=0}^{\infty} \|u_n\|_\infty < +\infty;$$

une telle série est dite *normalement convergente* sur A.

En appliquant les théorèmes sur les suites, on obtient les résultats suivants.

Théorème 1. Continuité de lu somme. Si, pour tout n, la fonction u,, est continue sur A et si la série converge uniformément sur **tout** compact de A vers f, alors f est continue sur A.

Théorème 2. Intégration terme à terme sur un intervalle compact. Supposons que, pour tout n, la fonction u_n est continue sur [a, b] et que la série converge uniformément sur [a, h]. Alors :

$$\int_a^b \sum_{n=0}^{\infty} u_n = \sum_{n=0}^{\infty} \int_a^b u_n.$$

Théorème 2 bis. Intégration terme à terme dans le cas non compact. Soit (u,) une suite de fonctions à valeurs complexes, intégrables sur I. On suppose :

$$\sum_{n=0}^{\infty} \int_I |u_n| < +\infty.$$

Alors la série $\Sigma u_n(t)$ converge sur I pour presque tout t, sa somme est intégrable sur 1 et on a :

$$\int_I \sum_{n=0}^{\infty} u_n = \sum_{n=0}^{\infty} \int_I u_n.$$

Théorème 3. Dérivation terme à terme. Supposons u_n de classe C^1 sur I et supposons que les séries Œu, et ΣDu_n convergent uniformément sur tout compact de I. Alors $\sum_{n=0}^{\infty} u_n$ est de classe C^1 et on a :

$$D\left(\sum_{n=0}^{\infty} u_n\right) = \sum_{n=0}^{\infty} Du_n.$$

Fonctions définies par des intégrales

On se donne une fonction $(x, t) \longmapsto f(x, t)$ définie sur A X I, à valeurs complexes, où A est un espace métrique et I un intervalle de R (ou, plus généralement, une partie

localement compacte de $\mathbf{R}^m$). On veut alors étudier la fonction :

$$x \mapsto F(x) = \int_I f(x,t)\,dt.$$

On dispose alors des trois résultats suivants. dont le premier est élémentaire.

Théorème 1. Dérivation sous le signe somme (cas des intervalles compacts). On suppose que A est un intervalle de R et $I = [a, b]$ un intervalle compact. Alors :

a) si f est continue sur A X [a, h], F est continue sur A :

b) si f admet une dérivée partielle $\partial f/\partial x$ continue sur A X [a, h], alors F est de classe C^1 sur A et :

$$DF(x) = \int_a^b \frac{\partial f}{\partial x}(x,t)\,dt.$$

Ce théorème a été utilisé par Leibniz.

Théorème 2. Dérivation sous le signe somme (Leibniz-Lebesgue). Soit A et 1 des intervalles de R. On fait les hypothèses suivantes :

a) pour tout t l'application $x \mapsto f(x, t)$ est continue sur A ;

b) pour tout $x \in$ A, l'application $t \mapsto f(x, t)$ est intégrable sur I ;

c) pour tout compact K de A, il existe $\varphi_K \geqslant 0$, intégrable sur I, telle que :

$$\forall t \in I,\ \forall x \in K,\quad |f(x,t)| \leqslant \varphi_K(t)$$

Alors f est continue sur A.

Si, de plus, on a :

d) f admet une dérivée partielle $\partial f/\partial x$ continue par rapport à x et satisfait à une hypothèse de domination sur tout compact K de A,

alors F est de classe C^1 sur A et on a :

$$DF(x) = \int_I \frac{\partial f}{\partial x}(x,t)\,dt.$$

Le problème de l'intégration de F se place dans le cadre plus général de la théorie des intégrales sur un espace produit, où on dispose du théorème suivant, qui permet le calcul d'une telle intégrale par intégrations successives.

Théorème de Fubini.

1. Soit $(x, y) \mapsto f(x, y)$ une fonction mesurable *positive* définie sur $\mathbf{R}^m$ X $\mathbf{R}^p$. Alors, il est équivalent de dire :

a) f'est intégrable sur $\mathbf{R}^m$ X $\mathbf{R}^p$;

b) pour presque tout x, la fonction $y \mapsto f(x, y)$ est intégrable sur $\mathbf{R}^p$ et la fonction :

$$x \mapsto \int_{\mathbf{R}^p} f(x,y)\,dy$$

est intégrable sur $\mathbf{R}^m$;

h') condition analogue à (6) en échangeant x et y.

Dans ces conditions, on a alors :

$$\iint_{\mathbf{R}^m \times \mathbf{R}^p} f(x,y)\,dx\,dy = \int_{\mathbf{R}^m}\left[\int_{\mathbf{R}^p} f(x,y)\,dy\right]dx = \int_{\mathbf{R}^p}\left[\int_{\mathbf{R}^m} f(x,y)\,dx\right]dy.$$

2. Si f est à valeurs complexes, (a) entraîne (b) et (h') et on a la formule de calcul de l'intégrale double ; mais ici les conditions (a), (h) et (b') ne sont plus équivalentes, comme le montrent des exemples classiques.

Emploi des distributions

Les théorèmes précédents sont satisfaisants en ce qui concerne la continuité et l'intégration mais le sont beaucoup moins en ce qui concerne la dérivation. Cela tient au fait que, dans les espaces classiques de fonctions, l'opérateur D de dérivation n'est pas continu. Un des avantages principaux de la théorie des distributions est précisément de fournir un

cadre théorique dans lequel la dérivation est une opération régulière au sens suivant : toute distribution T est indéfiniment dérivable et si une suite (T,) de distributions converge vers T, alors (DT,,) converge vers T (cf. DISTRIBUTIONS).

Le seul inconvénient est que la convergence des distributions ne peut pas être décrite par une norme. C'est pourquoi on introduit les *espaces de Sobolev : ce* sont des sous-espaces de $\mathcal{D}'$ qui présentent les mêmes avantages que $\mathcal{D}'$ mais qui, en outre, sont des espaces hilbertiens et sont bien adaptés à l'emploi de la transformation de Fourier et à l'étude des problèmes aux limites des équations aux dérivées partielles (cf. équations aux DÉRIVÉES PARTIELLES - Théorie linéaire).

Pour définir ces espaces, on considère l'espace vectoriel $\mathcal{S}'(\mathbf{R}^n)$ des distributions tempérées sur $\mathbf{R}^n$; on sait que la transformation de Fourier T $\mapsto \hat{T}$ est un automorphisme de S' (cf. DISTRIBUTIONS, chap. 4).

Pour tout s $\in$ R, on introduit le sous-espace vectoriel $H^s(\mathbf{R}^n)$ des distributions tempérées T telles que :

$$(1+|x|^2)^{s/2}\hat{T} \in L^2(\mathbf{R}^n).$$

Par l'isomorphisme T $\mapsto (1+|x|^2)^{s/2}\hat{T}$, on transporte à H' la structure hilbertienne de L^2, en posant :

$$(1)\quad (S|T) = \int_{\mathbf{R}^n} (1+|x|^2)^s \overline{\hat{S}(x)}\, \hat{T}(x)\, dx\,;$$

l'opérateur de dérivation :

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}$$

est alors une application *continue* de H^s dans $H^{s-|\alpha|}$, où $|\alpha| = \alpha_1 + \dots + \alpha_n$.

Lorsque $s = 0$, on a $H^0(\mathbf{R}^n) = L^2(\mathbf{R}^n)$. Plus généralement, pour $s = p \in N$, $H^p(\mathbf{R}^n)$ n'est autre que l'espace des distributions T telles que $D^\alpha T \in L^2(\mathbf{R}^n)$ pour tout α vérifiant $|\alpha| \leqslant p$; en outre, le produit scalaire (1) est alors équivalent au suivant :

$$(2)\quad ((S|T)) = \sum_{|\alpha|\leqslant p} \int_{\mathbf{R}^n} \overline{D^\alpha S(t)}\, D^\alpha T(t)\, dt.$$

Si maintenant $p = -s \in$ N, $H^{-p}(\mathbf{R}^n)$ est constitué des distributions T de la forme :

$$T = \sum_{|\alpha|\leqslant p}^{\infty} D^\alpha f_\alpha,\ f_\alpha \in L^2(\mathbf{R}^n).$$

En outre, on peut repasser de la théorie H^s à la théorie classique C^k grâce au théorème suivant.

Théorème de Sobolev. Si $k \in N$ et si $s > (n/2) + k$, alors $H^s(\mathbf{R}^n) \subset C^k(\mathbf{R}^n)$ et, si $f_n \to 0$ dans H^s, alors $D^\alpha f_n \to 0$ uniformément sur tout compact de $\mathbf{R}^n$ pour tout multi-indice α tel que $|\alpha| \leqslant k$.

Soit enfin K un compact de $\mathbf{R}^n$ et H^s_K le sous-espace vectoriel fermé de H^s constitué des distributions à support contenu dans K. On dispose alors du théorème de Rellich.

Théorème de Rellich. Si $s > t$, alors $H^s_K \subset H^t_K$ et, de toute suite bornée de H^s, on peut extraire une sous-suite convergente dans H'.

Une autre interprétation des espaces H^s est fournie par le théorème suivant.

Théorème. Soit $s \geqslant 0$ écrit sous la forme $s = p + \sigma$, où $0 \leqslant \sigma < 1$. Alors $f \in H^s(\mathbf{R}^n)$ si et seulement si $D^\alpha f \in L^2(\mathbf{R}^n)$ pour $|\alpha| \leqslant p$ et :

$$\iint_{\mathbf{R}^n \times \mathbf{R}^n} \frac{|D^\alpha f(x) - D^\alpha f(y)|^2}{|x-y|^{2\sigma+n}}\, dx\, dy < +\infty$$

si $|\alpha| = p$ et $\sigma > 0$.

Ce théorème permet notamment de prouver l'invariance de H^s par difféomor-

phisme et de définir l'espace H'(E) pour toute sous-variété C^∞ de dimension n.

Dans la théorie des équations aux dérivées partielles, on a besoin des espaces de Sobolev relatifs à un ouvert U borné de $\mathbf{R}^n$ dont la frontière X est une variété compacte de classe C^∞. On définit alors H'(Ü) comme l'ensemble des restrictions à U des éléments de $H^s(\mathbf{R}^n)$.

Les résultats concernant H^p et H^{-p} s'étendent à ce cas ainsi que le théorème de Rellich, où K est remplacé par $\bar{U}$. En outre, si $s > 1/2$, l'application $f \mapsto f_X$, définie sur $\mathcal{D}(\mathbf{R}^n)$, se prolonge de manière unique en une application linéaire continue de H'(Ü) dans $H^{s-1/2}(X)$, appelée *trace* sur X et notée $T \rightarrow T_{|X}$. Ces notions permettent de poser les problèmes aux limites de la théorie des équations aux dérivées partielles dans le cadre de la théorie des distributions (cf. équations aux DÉRIVÉES PARTIELLES Théorie linéaire).

7. Stabilité et consistance

On peut décrire les procédés *linéaires* d'approximation par le schéma général suivant : soit E et F des espaces vectoriels normés de fonctions, et u une application linéaire continue de E dans F. Un processus linéaire d'approximation de u est une suite (u,,) d'applications linéaires continues dc E dans F telles que, pour tout élément f de E. $u_n(f)$ converge vers $u(f)$.

Le cas le plus classique est celui où F = E et où u est l'application identique de E, c'est-à-dire où $u_n(f)$ converge vers f; nous en avons fourni de nombreux exemples aux chapitres 4 et 5, les plus significatifs étant ceux des séries de Fourier et de l'interpolation de Lagrange. Le cas où F = C, c'est-à-dire où u est une forme linéaire sur E est aussi très intéressant.

Stabilité

Un des problèmes les plus importants, surtout pour l'analyse numérique, concerne la *stabilité* du processus (u,,) : si l'on fait une petite erreur sur la fonction f, c'est-à-dire si on remplace f par une fonction g proche de f dans E, u,,(g) converge-t-elle vers un élément proche de $u(f)$? La réponse est fournie par le résultat suivant.

Théorème 1. Caractérisation des processus stables.

1. Si la suite $(\| u_n \|)$ des normes des applications linéaires u_n est bornée par un nombre M indépendant de n, alors, pour tout couple (f, g) d'éléments de E :

$$\|u_n(g) - u(f)\| \leqslant M\|g - f\| + \|u_n(f) - u(f)\|.$$

2. Réciproquement, si $(\| u_n \|)$ n'est pas bornée, alors, pour tout élément f de E, il existe une suite (g_n) d'éléments de E qui converge vers f et telle que $\| u_n(g_n) - u(f) \| \rightarrow +\infty$.

C'est pourquoi, on dit que le processus (u,,) est *stable* dans le cas (1) *et instable* dans le cas (2).

En fait, ce résultat n'est pertinent que si les espaces normés E et F sont complets, car, si E n'est pas complet, E est un sous-espace dense d'un espace normé plus gros, à savoir son complété Ê ; si bien que, en faisant une erreur sur f, on risque de passer à un élément g de Ê qui n'appartient pas à E. C'est le cas où, par exemple, Ê = $\mathcal{C}([a, b])$ muni de la norme de la convergence uniforme et si E $=$ $C^p([a, b])$. D'où l'intérêt du résultat suivant.

Théorème 1'. Stabilité et passage à un sous-espace vectoriel dense. Soit Ê et F des espaces de Banach et E un sous-espace dense de Ê. On considère une suite (u_n) d'applications linéaires continues de E

dans F satisfaisant aux deux conditions suivantes :

a) pour tout f de E, $u_n(f) \to u(f)$;

b) la suite des normes ($\| u_n \|$) est bornée par M.

Alors u est linéaire continue sur E, $\|u\| \leqslant$ M, et les applications u_n et u se prolongent de manière unique en des applications linéaires continues de $\hat{E}$ dans F, qui satisfont aux conditions (*a*) et (b) sur $\hat{E}$. Ainsi, le processus (u_n) est stable, non seulement sur E mais sur $\hat{E}$.

Dans le cas où E est *complet,* il est tout à fait remarquable que la convergence implique la stabilité.

Théorème 2 (stabilité et complétude). Soit E et F des espaces de Banach, (u_n) une suite d'applications linéaires continues de E dans F telle que, pour tout x de E, la suite ($u_n(x)$) converge dans F vers $u(x)$. Alors :

a) il existe M tel que $\| u_n \| \leqslant$ M ;

b) u est linéaire continue ;

c) $u_n \to u$ uniformément sur tout compact de E.

C'est le *théorème de Banach-Steinhaus* (cf. espaces vectoriels NORMÉS, chap. 4).

Le cas des séries de Fourier est à cet égard exemplaire ; c'est d'ailleurs un des exemples qui a été historiquement à l'origine du théorème général précédent. Ici E est l'espace vectoriel C(T) des fonctions continues l-périodiques muni de la norme N_∞ ; cet espace est complet ; u_n est la forme linéaire qui, à tout élément f de C(T), associe la valeur de la somme partielle $s_n(f)$ de sa série de Fourier en un point donné x ; $u_n(f)$ n'est autre que $D_n * f(x)$, où D,, est le noyau de Dirichlet
On montre que :

$$\|s_n\|_\infty = \|D_n\|_1 = \int_0^1 \left|\frac{\sin(2n+1)\pi t}{\sin \pi t}\right| dt \sim \frac{2}{\pi} \ln n.$$

Ainsi, $\| D_n \|_1$ n'est pas borné ; il existe donc des fonctions continues dont la série de Fourier diverge au point x. Cependant, la convergence uniforme de $s_n(f)$ vers f est assurée par exemple si f appartient au sous-ensemble dense des fonctions périodiques de classe C^1 (cf. SÉRIES TRIGONOMÉTRIQUES).

Comme nous l'avons déjà vu au chapitre 5, un phénomène analogue se produit pour les processus interpolatoires.

En revanche, le processus d'approximation par les sommes partielles $s_n(f)$ est stable sur (C(T), N_2), et même sur (L'(T), N_2) car ici s_n est un projecteur orthogonal et $\| s_n \|_2$ = 1. De même, si on prend les sommes de Fejer $\sigma_n(f) = F_n * f$, alors $\| \sigma_n \|_\infty = 1$, ce qui assure la stabilité sur ($\mathcal{C}$(T), N_∞).

On observera que, dans ce dernier cas, le noyau est positif. Cette condition assure la stabilité de manière très générale.

Théorème 3. Continuité des opérateurs linéaires positifs. Soit u un endomorphisme de l'espace vectoriel C(T), ou $\mathcal{C}$([a, h]), muni de la norme N_∞. On suppose u *positif;* c'est-à-dire $u(f) \geqslant 0$ pour tout $f \geqslant 0$. Alors u est continu ; plus précisément :

$$N_\infty(u(f)) \leqslant N_\infty(u(1)) N_\infty(f).$$

Autrement dit, $\|u\| \leqslant N_\infty(u(1))$.

Si maintenant (u_n) est une suite d'opérateurs linéaires positifs qui converge simplement vers un opérateur u, auquel cas u est positif, alors la condition de stabilité est satisfaite pour u_n :

$$\| u_n \| \leqslant \sup_n N_\infty(u_n(1)).$$

En outre, le théorème suivant assure la convergence du processus sous des hypothèses très faibles.

Théorème 4 (théorème de Korovkine). Soit (u,,) une suite d'endomorphismes posi-

tifs de $\mathcal{C}([a, b])$ muni de la norme N_∞ et u un endomorphisme de cet espace satisfaisant aux conditions suivantes : $u_n(f) \to u(f)$ uniformément sur $[a, b]$ pour chacune des trois fonctions $x \mapsto 1$, $x \mapsto x$ et $x \mapsto x^2$. Alors $u_n(f) \to u(f)$ pour toute fonction continue f.

Ce théorème fournit une nouvelle démonstration, due à Bernstein, du théorème d'approximation polynomiale de Weierstrass.

Théorème 5. On considère les polynômes de Bernstein :

$$B_{n,k} = C_n^k x^k (1-x)^{n-k}, \quad 0 \leqslant k \leqslant n.$$

Si à tout $f \in \mathcal{C}([a, b])$ on associe la fonction polynomiale $u_n(f)$ définie par :

$$(u_n(f))(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) B_{n,k}(x),$$

alors $u_n(f)$ tend vers f uniformément sur $[a, b]$.

La convergence n'est pas très rapide car, si f est, par exemple, la fonction $x \mapsto x^2$, on a $\| u_n(f) - f \| = 1/4n$.

D'ailleurs, l'approximation polynomiale par des opérateurs positifs u_n ne peut *jamais* être rapide, car on démontre que $\| u_n(f) - f \|_\infty$ n'est pas négligeable devant $1/n^2$ pour l'une au moins des trois fonctions $x \mapsto 1$, $x \mapsto x$, $x \mapsto x^2$.

Enfin, le théorème de Korovkine s'étend aussitôt au cas des fonctions continues périodiques; il suffit que la convergence de $u_n(f)$ vers f soit assurée pour les trois fonctions $x \mapsto 1$, $x \mapsto \cos 2\pi x$, $x \mapsto \sin 2\pi x$. On retrouve ainsi la convergence uniforme des sommes de Fejer $\sigma_n(f)$ pour toute fonction continue périodique ; ici encore la convergence est *lente* car $\| \sigma_n(f) - f \|_\infty$ est exactement de l'ordre de $1/n$ lorsque $f(x) = |\cos 2\pi x|$.

Consistance

Dans de nombreux exemples, le processus d'approximation (u_n) d'une application u de E dans F se présente sous la forme suivante : on se donne une **suite** (E_n) de sous-espaces vectoriels de dimension finie d'un espace vectoriel normé E de fonctions telle que

$$\bigcup_{n=0}^{\infty} E_n$$

soit dense dans E. On dit alors que le processus (u_n) est *consistant* si, pour tout élément $\varphi \in E_n$, $u_n(\varphi) = u(\varphi)$, c'est-à-dire $u_n = u$ sur E_n.

Un cas particulièrement important est celui où u_n est défini à partir d'une suite (p_n) de *projecteurs* de E sur E_n grâce à la relation $u_n = u \circ p_n$. Si $E = F = C(T)$ muni de la norme N_2 et si $E_n = \mathcal{T}_n$ est l'espace vectoriel des polynômes trigonométriques de degré $\leqslant n$, on peut prendre pour p_n le projecteur orthogonal de E sur E_n. Il en est de même si $E = F = \mathcal{C}([a, b])$, muni de la norme :

$$f \mapsto N_{2,\pi}(f) = \int_a^b |f(t)|^2 \pi(t)\, dt,$$

où π est un *poids,* c'est-à-dire une fonction positive intégrable donnée sur $[a, b]$ et si $E_n = \mathcal{P}_n$.

Lorsque $E = F = \mathcal{C}([a, b])$ muni de la norme N_∞ et $E_n = \mathcal{P}_n$, les interpolations de Lagrange ou, plus généralement, d'Hermite constituent aussi des exemples significatifs.

Pour étudier la convergence des processus consistants, l'idée est de comparer $u_n(f)$ à un élément $\varphi_n(f)$ approchant f le mieux possible. Plus précisément, on note $\delta_n(f)$ la distance d'un élément f de E à E_n, définie par :

$$\delta_n(f) = d(f, E_n) = \inf_{\varphi \in E_n} \| f - \varphi \|.$$

Comme E,, est de dimension finie, il existe au moins un élément $\varphi_n \in E_n$ tel que $\|f - \varphi_n\| = d(f, E_n)$; on dit alors que φ_n optimise l'approximation de f par les éléments de E,,. L'étude de ce problème est abordée au chapitre 8.

Dans ces conditions, comme $u_n(\varphi_n) = u(\varphi_n)$,

$$\|u(f) - u_n(f)\| = \|(u - u_n)(f - \varphi_n)\| \leqslant \delta_n(f)(\|u\| + \|u_n\|).$$

Le problème de la convergence du processus (u,,) et de sa rapidité de convergence est alors séparé en deux questions totalement distinctes :
l'évaluation de $\delta_n(f)$ qui est liée à la régularité de la fonction que l'on veut approcher ;
l'évaluation de $\| u_n \|$ qui mesure la qualité du processus d'approximation.

8. Optimisation de l'approximation ; rapidité de convergence

Optimisation de l'approximation

Avec les notations du chapitre précédent, nous allons étudier les deux problèmes suivants :

a) l'unicité de l'élément φ_n de E,, optimisant l'approximation de f par les éléments de E,, ; il est alors intéressant de construire des méthodes explicites de calcul de φ_n ;

b) la distance $\delta_n(f)$ tend-elle vers 0 si n tend vers $+\infty$? Si oui, déterminer la vitesse de convergence en fonction des propriétés de f.

Voici deux exemples classiques.

Dans le premier, E est un espace vectoriel de fonctions l-périodiques à valeurs complexes et E,, est le sous-espace vectoriel $\mathcal{T}_n$ des polynômes trigonométriques de degré $\leqslant n$, c'est-à-dire des combinaisons linéaires des fonctions exponentielles $t \mapsto e^{2i\pi pt}$, où $p \leqslant n$.

Dans le second, E est un espace vectoriel de fonctions définies sur $[a, b]$ et E_n est le sous-espace vectoriel $\mathcal{P}_n$ des polynômes de degré $\leqslant$ n.

Bien entendu, les réponses aux problèmes précédents vont dépendre du type de convergence considéré. Nous examinerons principalement le cas des normes N_2 (approximation en moyenne quadratique) et N_∞ (approximation uniforme).

Unicité de φ_n

Théorème 1. L'unicité de φ_n est assurée lorsque la boule unité est *strictement convexe*, c'est-à-dire si les relations $\|x\| = \|y\| = 1$ et $\alpha x + \beta y = 1$ avec $\alpha > 0$, $\beta > 0$, ci $+ \beta = 1$ impliquent $x = y$.

Cette dernière condition est réalisée pour l'espace $E = L^2(\mu)$ des fonctions de carré intégrable pour une mesure μ muni de la norme N_2, et, plus généralement, pour les espaces $L^p(\mu)$ pour $1 < p < +\infty$. Elle ne l'est pas pour l'espace $L^1(\mu)$, ni pour l'espace $E = \mathcal{C}([a, h])$ muni de la norme N_∞ ; on peut relier ce phénomène à la forme des boules de $\mathbf{R}^2$ pour ces mêmes normes (cf. figure in espaces vectoriels NORMÉS). D'ailleurs, dans ces deux cas, on peut donner des exemples où il n'y a pas unicité de φ_n : il suffit par exemple de prendre $E = \mathcal{C}([0, 1])$ muni de la norme N,. Si E_0 est la droite vectorielle engendrée par la fonction $t \mapsto t$, alors $d(1, E_0)$ est atteinte pour toutes les fonctions $\varphi(t) = \alpha t$ où $0 \leqslant \alpha \leqslant 2$. 11 en est de même si E est l'espace C(T) des fonctions continues l-périodiques et si E_0 est la droite
vectorielle engendrée par la fonction $t \mapsto \sin 2\pi t$. Cela tient au fait que toutes les

fonctions de E_0 s'annulent en un même point.

Ainsi, le problème de l'unicité de la meilleure approximation uniforme dans $E = \mathcal{C}([a, h])$ est assez délicat. Pour le résoudre, on introduit le concept de sous-espace vectoriel *régulier* de fonctions (cf. *supra*, chap. 5).

On a alors le théorème suivant, dû à Haar.

Théorème 2. Soit E_n un sous-espace vectoriel de $\mathcal{C}([a, h])$, ou de l'espace vectoriel C(T) des fonctions continues l-périodiques, muni de la norme de la convergence uniforme. Il est équivalent de dire :

a) E,, est régulier ;

b) l'unicité de φ_n est assurée pour tout élémentf'de $\mathcal{C}([a, h])$ (ou de C(T)).

Ce théorème s'applique à la meilleure approximation polynomiale uniforme des fonctions continues, car le sous-espace vectoriel des fonctions polynômes de degré $\leqslant n$ est régulier. Il s'applique aussi à la meilleure approximation uniforme des fonctions continues périodiques par les polynômes trigonométriques, car le sous-espace vectoriel des polynômes trigonométriques de degré $\leqslant n$ est lui aussi régulier.

Caractérisation et calcul explicite de φ_n

Théorème 3 (cas des normes quadratiques). Soit E,, un sous-espace vectoriel de dimension finie d'un espace vectoriel hilbertien E. Alors φ_n n'est autre que la projection orthogonale de f sur E,, ; autrement dit, $\varphi_n = p_n(f)$, où p_n est le projecteur orthogonal sur E,,. En particulier, l'application $f \mapsto \varphi_n$ est linéaire continue et :

$$N_2(\varphi_n) \leqslant N_2(f),$$
$$\delta_n(f)^2 = N_2(f-\varphi_n)^2 = N_2(f)^2 - N_2(\varphi_n)^2.$$

Si on se donne une base $(e_0, \dots e_n)$ de E_n, on peut calculer explicitement φ_n. Si cette base est orthonormée, on a :

$$\varphi_n = \sum_{j=0}^{n} (e_j | f) e_j,$$

$$\delta_n(f)^2 = N_2(f)^2 - \sum_{j=0}^{n} |(e_j|f)|^2.$$

Sinon, les composantes de φ_n se calculent à l'aide des déterminants de Gram, en particulier :

$$\delta_n(f)^2 = \frac{\text{Gram}(f, e_0, \dots, e_n)}{\text{Gram}(e_0, \dots, e_n)},$$

où $\text{Gram}(x_1, \dots x_p) = \det((x_i \; x_j))$ (cf. espace de HILBERT).

Le cas des normes uniformes est plus délicat. Pour simplifier, nous nous limiterons au cas de l'approximation polynomiale uniforme des fonctions continues à valeurs réelles, c'est-à-dire $E = \mathcal{C}([a, b], \mathbb{R})$ et E,, $= \mathcal{P}_n$, espace des polynômes de degré $\leqslant$ II. On dit qu'une suite strictement croissante $(t_1, \dots, t_r)$ de points de $[a, b]$ est *alternante* pour une fonction F de $\mathcal{C}([a, h])$ si, pour toutj $\in [1, r]$, $F(t_j) = N_\infty(f)$ et si, pour tout $j \in [1, r-1]$, $F(t_j)F(t_{j+1}) < 0$; r s'appelle la longueur de la suite. On a alors le théorème suivant.

Théorème 4. Théorème d'équioscillation de Tchebychev. Soit $f \in \mathcal{C}([a, b])$ et φ une fonction polynomiale de degré $\leqslant n$. Il y a équivalence entre :

a) la fonction φ est égale au polynôme de meilleure approximation φ_n ;

b) il existe une suite alternante relativement à $f - \varphi$ de longueur au moins $n + 2$.

Ce théorème détermine φ_n, mais il n'existe aucun algorithme simple dès que n est un peu grand. C'est néanmoins possible lorsque f est une fonction poly-

nomiale de degré $n + 1$. Grâce au théorème précédent, on démontre en effet la propriété extrémale des polynômes de Tchebychev, avec ici $[a, b] = [-1, 1]$ (cf. *supra,* chap. 5).

Théorème 5. Soit T_n le polynôme de Tchebychev de degré n ; alors :

a) $N_\infty(T_n) = 2^{n-1}$ et. pour tout polynôme unitaire P de degré n, on a $N_\infty(P) \geqslant 1/2^{n-1}$, avec égalité si et seulement si $P = T_n/2^{n-1}$;

b) soit P un polynôme de degré $n+1$, de coefficient dominant a_{n+1}, que l'on écrit :

$$P = \sum_{p=0}^{n+1} \lambda_p T_p ;$$

alors :

$$d(P, E_n) = |\lambda_{n+1}| = \frac{1}{2^n} |a_{n+1}|,$$

$$\varphi_n = \sum_{p=0}^{n} \lambda_p T_p .$$

On notera que l'application $P \mapsto \varphi_n$ est linéaire, mais, malheureusement, il n'est pas vrai, contrairement au cas des normes N_2, que, si P est de degré $n+2$, φ_n s'obtienne en prenant d'abord $Q = \varphi_{n+1}(P)$ et ensuite $\varphi_n(Q)$. Plus généralement, l'application $f \mapsto \varphi_n$ n'est pas linéaire. Cependant, en pratique, on utilise souvent une méthode *télescopique :* on commence par approcher f par un polynôme P_{n+p} de degré $n+p$ suffisamment élevé pour obtenir la précision cherchée, par exemple en utilisant le développement taylorien de f à l'origine (s'il est connu), puis on développe P_{n+p} dans la base des polynômes de Tchebychev et on tronque le développement à l'ordre n ; on obtient ainsi un polynôme ψ_n, qui n'est pas égal à φ_n, mais dont on démontre qu'il en diffère fort peu. Une variante de cette méthode dans le cas périodique consiste à développer la fonction $x \mapsto f(\cos x)$ en série de Fourier :

$$f(\cos x) = \sum_{p=0}^{\infty} A_p \cos px$$

(ce qui revient à développer f en série de polynômes de Tchebychev :

$$f(t) = \sum_{p=0}^{\infty} \lambda_p T_p(t),$$

et à effectuer le changement de variable $t = \cos x$). On calcule les coefficients de Fourier A_p et le polynôme :

$$\psi_n(t) = \sum_{p=0} \lambda_p T_p(t)$$

fournit une excellente approximation de f pour un polynôme de degré $< n$. On peut prouver que le gain par rapport au développement taylorien est de $1/2^n$, ce qui est en accord avec les résultats du chapitre 5.

Convergence de $\delta_n(f)$

Il nous reste maintenant à étudier le second problème, qui concerne la convergence vers 0 de $\delta_n(f)$. Tout d'abord, le théorème d'approximation polynomiale uniforme de Weierstrass signifie que $\delta_n(f)$ tend vers 0 dans les deux cas classiques $E = C^0([a, b])$, $E_n = \mathcal{P}_n$ et $E = C(T)$, $E_n = \mathcal{T}_n$ pour la norme uniforme N_∞.

Comme la norme N_∞ est plus fine que N_2 sur un intervalle compact, le même résultat est valable dans ce cas pour la norme N_2 et, par suite, pour les espaces hilbertiens $L^2([a, b])$ et $L^2(T)$.

Les résultats explicites fournis par le théorème 3 permettent d'étudier des cas plus fins.

Théorème 6 (théorème de Muntz). Soit E = [0, 1], (α_n) une suite strictement croissante de nombres réels positifs telle que $\alpha_0 = 0$, et E_n l'espace vectoriel engendré par $e_{\alpha_0}, \dots, e_{\alpha_n}$, où $e_\beta(t) = t^\beta$.

1. Si E est muni de la norme N_2, il est équivalent de dire :

a) pour tout $f \in \mathcal{C}([0, 1])$, $\delta_n(f) \to 0$, c'est-à-dire la suite (e_{α_n}) est totale ;

b) la série

$$\sum_{n=1}^{\infty} \frac{1}{\alpha_n}$$

diverge.

2. Cette équivalence reste valable si E est muni de la norme N_∞.

Pour démontrer ce résultat, on calcule la distance de e_β à E_n par la méthode de Gramm indiquée plus haut et on montre que cette distance tend vers 0 si et seulement si la condition *b* est réalisée.

La méthode utilisée pour passer de N_2 à N_∞ consiste en une inégalité du type Sobolev : si f est nulle en 0 et de classe C^1, on a :

$$N_\infty(f) \leqslant \sqrt{b-a} \times N_2(Df) ;$$

pour approcher uniformément f par des éléments de E_n, il suffit donc d'approcher Df en moyenne quadratique. Ainsi, le cas de N_∞, d'accès difficile par une méthode directe, est résolu grâce à un détour par N_2. Ces méthodes sont d'usage constant en analyse fonctionnelle.

Enfin, la rapidité de convergence de $\delta_n(f)$ vers 0 dépend de la *régularité* de f. Ce phénomène est analogue à celui des séries de Fourier (cf. SÉRIES TRIGONOMÉTRIQUES). Pour mesurer de façon précise la régularité des fonctions continues, on introduit le *module de continuité*. Soit f une fonction continue sur un intervalle I de R ; pour tout $\delta > 0$, on pose :

$$\omega_f(\delta) = \sup_{|x-y| \leqslant \delta} |f(x) - f(y)|, \quad \text{pour } x, y \in I ;$$

cette notion est très commode car elle permet de décrire différentes classes de fonctions :

la fonction f est uniformément continue sur I si et seulement si $\omega_f(\delta) \to 0$ si $\delta \to 0$;

la fonction f est lipschitzienne de rapport k sur 1 si et seulement si $\omega_f(\delta) \leqslant k\delta$.

Plus généralement, f est holdérienne d'ordre α, pour $0 < \alpha \leqslant 1$, si et seulement si $\omega_f(\delta) \leqslant k\delta^\alpha$.

Dans le cas des normes N_∞, on introduit les classes de fonctions suivantes :

l'espace vectoriel $\text{Lip}_\alpha([a, b])$ des fonctions holdériennes sur $[a, b]$ d'ordre α, $0 < \alpha \leqslant 1$, muni de la norme :

$$f \mapsto N_\infty(f) + \lambda_\alpha(f),$$

où :

$$\lambda_\alpha(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|} ;$$

– pour tout $p \geqslant 0$, l'espace vectoriel $\text{Lip}_{\alpha,p}([a, b])$ des fonctions de classe C^p sur [a, b] telles que $D^p f \in \text{Lip}_\alpha([a, b])$ muni de la norme :

$$f \mapsto N_\infty(f) + \dots + N_\infty(D^p f) + \lambda_\alpha(D^p f).$$

Le cas des fonctions périodiques est analogue. Les espaces vectoriels normés $\text{Lip}_{\alpha,p}$ sont complets.

Pour étudier $S_n(f)$, on commence par le cas des fonctions périodiques et on en déduit le cas $\mathcal{C}([-1, 1])$ par le changement de variable $x = \cos \pi t$. Considérons donc une fonction continue 1-périodique f. L'idée est d'approcher f par convolution avec des polynômes trigonométriques χ_n constituant une approximation

de la mesure de Dirac et d'évaluer $\|f - f * \chi_n\|_\infty$.

Nous connaissons déjà les noyaux de Dirichlet et de Fejer :

$$D_n(f) = \frac{\sin(2n+1)\pi t}{\sin \pi t}$$
$$F_n(t) = \frac{1}{n}\left(\frac{\sin n\pi t}{\sin \pi t}\right)^2$$

Mais le premier, n'étant pas positif, ne convient pas pour toutes les fonctions continues, et le second, bien que positif, fournit des convergences très lentes, même si f est très régulière. On utilise ici le noyau de Jackson :

$$J_n(t) = \frac{1}{a_n}\left(\frac{\sin n\pi t}{\sin \pi t}\right)^4,$$

où a_n est une constante de normalisation. On a alors le théorème suivant.

Théorème 7 (Jackson). Il existe une constante $\beta > 0$ telle que, pour toute f continue périodique, on ait :

$$N_\infty(f - f * J_n) \leqslant \beta \omega_f\left(\frac{1}{n}\right)$$

En particulier :

$$\delta_n(f) \leqslant \beta\, \omega_f\left(\frac{1}{n}\right)$$

Ainsi, plus f est régulière, plus la décroissance de $\delta_n(f)$ est rapide. En particulier, si $f \in \mathrm{Lip}_\alpha$, $\delta_n(f) = 0(1/n^\alpha)$.

Pour $f \in \mathrm{Lip}_{\alpha,p}$, on introduit les noyaux :

$$J_{n,r}(t) = \frac{1}{a_{n,r}}\left(\frac{\sin n\pi t}{\sin \pi t}\right)^{2r},$$

et le théorème 7 se généralise par le théorème suivant.

Théorème 7'. Il existe $\beta > 0$ tel que, pour toute fonction périodique de classe C^p, on ait :

$$\delta_n(f) \leqslant \frac{1}{n^p}\beta\, \omega_{D^p f}\left(\frac{1}{n}\right).$$

En particulier, si on a $f \in \mathrm{Lip}_{\alpha,p}$, alors $\delta_n(f) = O(1/n^{p+\alpha})$.

Il est tout à fait remarquable que, inversement, si f est « bien approchée » par les polynômes trigonométriques, c'est-à-dire si $\delta_n(f)$ tend vers 0 assez rapidement, alors f est assez régulière.

Théorème 8 (Bernstein). Si $0 < \alpha < 1$, $p \geqslant 0$, alors :

$$\delta_n(f) = O\left(\frac{1}{n^{p+\alpha}}\right) \Rightarrow f \in \mathrm{Lip}_{\alpha,p}.$$

Cela prouve que les estimations fournies par les théorèmes 7 et 7' sont les meilleures possibles. Enfin, si f est C^∞ (resp. holomorphe dans une bande $|\mathrm{Im}\, z| < y$), alors $\delta_n(f)$ est à décroissance rapide (resp. exponentielle) ; ici encore, les réciproques sont exactes. On notera enfin qu'il existe des fonctions continues f telles que $\delta_n(f)$ décroisse arbitrairement lentement. Enfin, les théorèmes 7 et 7' s'appliquent sans changement au cas des fonctions continues sur un intervalle $[a, b]$, mais ici, les réciproques sont plus délicates.

Nous ne développerons pas non plus le cas, étudié par Favard et Krein, où $\mathcal{C}([a, b])$ est muni de la norme quadratique ; les théorèmes 7 et 7' s'étendent sans changement, les espaces $\mathrm{Lip}_{\alpha,p}$ étant remplacés par les espaces de Sobolev $H^{\alpha+p}$.

Convergence des processus d'approximation consistants

Nous nous bornerons ici à étudier le cas $E = C(T)$, ou $E = \mathcal{C}([a, b])$, où $u = I_E$ et où le processus (u_n) d'approximation est défini par une suite de projecteurs p_n de E sur E_n. D'après les résultats du chapitre 7, la convergence de $p_n(f)$ vers un élément f de E est contrôlée par la relation :

(1) $\delta_n(f) \leqslant \|p_n(f) - f\| \leqslant \delta_n(f)(1 + \|p_n\|)$.

Ayant étudié précédemment la rapidité de convergence vers 0 de $\delta_n(f)$, il nous reste à examiner le comportement de $\|p_n\|$. Voici tout d'abord un cas idéal.

Théorème 1. Processus consistants et stables. Tout processus à la fois consistant et stable est convergent sur E tout entier car, dans ces conditions, $\delta_n(f) \to 0$ et $\|u_n\| \leqslant M$. En outre, la vitesse de convergence est de même ordre de grandeur que celle du processus optimal.

Nous examinerons maintenant d'abord la possibilité d'obtention de processus d'approximation *polynomiale à* la fois stables et consistants. Ici encore, la situation est radicalement différente suivant qu'il s'agit d'approximation en moyenne quadratique ou d'approximation uniforme.

Approximation polynomiale quadratique

Théorème 2. Stabilité et convergence de l'approximation polynomiale en moyenne quadratique.

1. *Cas des fonctions périodiques.* E est l'espace vectoriel $L^2(T)$ des fonctions de carré intégrable l-périodiques muni de la norme N_2, $E_n = \mathcal{C}_n$ est le sous-espace vectoriel des polynômes trigonométriques de degré $\leqslant n$ et p_n est le projecteur orthogonal de E sur E,, qui, à toute fonction f, associe la somme partielle $s_n(f)$ de sa série de Fourier.

2. *Cas des fonctions définies sur un intervalle* I *de* R. On se donne un poids π défini sur 1, et E est l'espace vectoriel $L^2(I, \pi)$ des fonctions de carré intégrable pour la mesure de densité π muni de la norme $N_{2,\pi}$; ici $E_n = \mathcal{P}_n$, sous-espace vectoriel des polynômes de degré $\leqslant n$ etp,, est le projecteur de E sur E,, qui associe à f la somme partielle $s_n(f)$ de son développement suivant le système de polynômes orthogonaux défini par π.

Dans les deux cas, $\|p_n\| = 1$, donc le processus est à la fois stable et consistant ; nous avons vu précédemment qu'il est d'ailleurs optimal. En particulier, la convergence a lieu pour tout élément f de E et elle est d'autant plus rapide que f est *régulière :* Il existe $\beta > 0$ tel que :

$$f \in H^{\alpha+p} \Rightarrow \|f - s_n(f)\|_2 \leqslant \beta \frac{1}{n^{p+\alpha}}.$$

Approximation polynomiale uniforme

Théorème 3 (théorème de Bernstein-Fuber). Convergence de l'approximation polynomiale trigonométrique uniforme. On prend $E = C(T)$ muni de la norme N_∞ et $E_n = \mathcal{C}_n$.

1. Le processus de Fourier n'est pas stable, car la norme du projecteur p_n : $f \mapsto s_n(f)$ est égale à $\|D_n\|_1$, où D,, est le noyau de Dirichlet, et $\|D_n\|_1 \sim (2/\pi)\ln n$. Cependant, la convergence uniforme de $s_n(f)$ vers f a bien lieu dès que le module de continuité $\omega_f(1/n)$ est négligeable devant $1/\ln n$, et, en particulier, si f est holdérienne d'ordre α. En outre, la perte de rapidité par rapport au processus optimal est faible, puisque :

$$\|f - s_n(f)\|_\infty \leqslant \delta_n(f)\left(1 + \frac{2}{\pi}\ln n\right)$$

En particulier, si $f \in \mathrm{Lip}_{\alpha,p}$:

$$\|f - s_n(f)\|_\infty = O\left(\frac{\ln n}{n^{p+\alpha}}\right);$$

ainsi, la convergence est d'autant plus rapide que f est régulière.

2. Le processus de Fourier est optimal *parmi les processus consistants.* On peut en effet démontrer que, pour tout projecteur continu p_n de $E = C(T)$ sur $\mathcal{C}_n$, on a $\|p_n\| \geqslant \|D_n\|_1$.

3. En particulier, le processus consistant (p_n) n'est *jamais* stable, et, comme

(E, N,) est complet, il existe des fonctions continues f telles que $\|p_n(f)\|_\infty$ soit non borné ; bien entendu, pour une telle fonction f, la suite $p_n(f)$ ne tend pas vers f.

Ainsi, il n'existe pas de processus d'approximation polynomiale trigonométrique à la fois consistant et uniforme, ou, ce qui revient au même, convergent sur tout C(T). A fortiori, il n'existe pas de processus consistant positif, ce qui résulte aussi de la méthode de Korovkine.

Par changement de variable, on obtient le théorème suivant.

Théorème 3'. Convergence de l'approximation polynomiale uniforme.

On prend E = $\mathcal{C}([a, b])$ muni de la norme uniforme et E,, = $\mathcal{P}_n$.

1. Le processus de développement de f en série de polynômes de Tchebychev n'est pas stable, car $\|p_n\| \sim 1/\ln n$. Mais il converge pour les fonctions telles que $\omega_f(1/n)$ soit négligeable devant $1/\ln n$, et d'autant plus rapidement que f est plus régulière.

2. Ce processus est optimal parmi les processus consistants.

3. En particulier, il n'existe pas de procédé d'approximation polynomiale uniforme à la fois consistant et stable. On notera que (("([a. h]), N,) étant complet, la convergence simple n'est même pas assurée.

En outre. pour les fonctions suffisamment régulières, on peut remonter de l'approximation en moyenne quadratique à l'approximation uniforme. Il suffit pour cela de prouver que la série $\Sigma\alpha_n(f)e_n$ donnant le développement de f converge normalement sur l'intervalle considéré. En effet, dans ces conditions, $s_n(f)$ converge uniformément vers une fonction continue g ; comme la convergence uniforme implique la convergence en moyenne quadratique sur un intervalle compact, $s_n(f)$ converge aussi vers g en moyenne quadratique, et. par unicité de la limite, g $= f$. Pour établir la convergence normale, on majore $\| e_n \|_\infty$ et on utilise le fait que, plus f est régulière, plus la suite $(\alpha_n(f))$ décroît rapidement.

Le cas des séries de Fourier est bien classique : sif est continue et de classe C^1 par morceaux, la série :

$$\sum_{n=-\infty}^{+\infty} |\alpha_n(f)|$$

est convergente, ce qui assure la convergence normale de la série considérée puisque $\|e_n\|_\infty = 1$ (cf. SÉRIES TRIGONOMÉTRIQUES). On dispose de résultats analogues pour les développements en séries de polynômes orthogonaux (cf. polynômes ORTHOGONAUX). Dans le cas des développements en série de polynômes de Legendre, si f est continue et de classe C^1 par morceaux sur [- 1, + 1], la série $\Sigma\ \alpha_n(f)\ e_n$ converge normalement sur tout compact de]– 1, 1[et sifest de classe C^2, cette convergence est normale sur [- 1, + 1]. Pour les développements en série de polynômes de Hermite, si f est continue et de classe C^1 par morceaux sur R et si f etf" sont de carré intégrable pour le poids π, alors la convergence est normale sur tout compact de R.

Cette même idée s'applique encore à l'étude de la convergence uniforme des développements en série de fonctions propres des équations de Sturm-Liouville (cf. équations DIFFÉRENTIELLES, chap. 3) ou des équations intégrales de Fredholm (cf. équations INTÉGRALES).

Dans le même ordre d'idée, signalons le théorème de Rademacher-Menchov : si $\Sigma\ |\alpha_n|^2(\ln n)^2$ est fini, alors $\Sigma\ \alpha_n e_n$ converge vers f presque partout.

Nous examinerons enfin le cas des processus interpolatoires.

Théorème 4. Convergence des processus interpolutoires. Ici E = ?([a, h]) muni de N_∞. Soit S un système interpolateur sur [a, b] et L,, le projecteur de E sur $\mathcal{T}_n$ qui à tout élément f de E associe le polynôme d'interpolation de Lagrange (ou de Hermite) L,(f) (cf. chap. 5).

1. Pour tout S, $\|L_n\|$ domine $\ln n$; les processus interpolatoires ne sont donc ni stables ni convergents sur E.

2. Si on prend pour S le système de Tchebychev, où $[a, b] = [-1, 1]$ et où **a, =** $\cos[(2j+1)\pi/2n]$ pour $0 \leqslant j \leqslant n$ 1, et $L_n(f)$ le polynôme d'interpolation de Lagrange associé à f, alors $\|L_n\| \sim \ln n$. Autrement dit, l'interpolation de Tchebychev est sensiblement optimale parmi les processus consistants d'approximation polynomiale. En particulier, $L_n(f)$ converge uniformément vers f dès que $\omega_f(1/n)$ est négligeable devant $1/\ln n$ et, si $f \in \mathrm{Lip}_{\alpha,p}$:

$$\|L_n(f)-f\|_\infty = O\left(\frac{1}{n^{p+\alpha}}\right).$$

En revanche, si on prend des subdivisions à pas constant, $\|L_n\|$ croît exponentiellement, si bien que le processus peut fort bien diverger, même sif est analytique.

Généralisations

Comme l'approximation polynomiale uniforme ne fonctionne pas bien, on est amené à choisir des fonctions un peu plus générales que les polynômes, à savoir polynomiales par morceaux.

Le cas le plus simple est celui des fonctions affines par morceaux.

Théorème 5. Ici E = $\mathcal{C}([a, h])$ muni de N,. Soit S_n la subdivision de [a, b] à pas constant $(b-a)/n$ et p_n le projecteur de E sur le sous-espace vectoriel A, des fonctions continues et affines sur chaque intervalle $[t_j, t_{j+1}]$ de S,,. Alors le processus consistant (p_n) est stable car $\|p_n\| = 1$. En outre :

$$\|f-p_n(f)\|_\infty \leqslant \omega_f\left(\frac{b-a}{n}\right);$$

mais la rapidité de convergence n'est guère améliorée si f est régulière (cf. *supra*, chap. 5).

C'est pourquoi on peut alors recourir aux fonctions spline.

Théorème 5'. Convergence de l'approximation pur des fonctions spline. Soit encore S,, comme dans le théorème 5 et soit A,,,, l'espace vectoriel des fonctions spline cubiques associées à S_n et p_n le projecteur de E sur A,,, qui à toutf $\in$ E associe la fonction spline cubique $S_n(f)$ interpolant f (cf. *supra,* chap. 5). Alors le processus consistant (p_n) est stable et :

$$\|f-S_n(f)\|_\infty \leqslant \beta\,\omega_f\left(\frac{b-a}{n}\right)$$

Cette fois, la rapidité de convergence est améliorée si f est plus régulière (jusqu'à la classe C^4).

Le cas des fonctions périodiques est entièrement analogue.

JEAN-LOUIS OVAERT et JEAN-LUC VERLEY

Bibliographie

* Représentation des fonctions

J. DIEUDONNÉ, *Calcul infinitésimal*, Hermann, 2' éd. 1980 / A. KOLMOGOROV & S. FOMINE, *Éléments de lu théorie des fonctions et de l'analyse fonctionnelle*, M.I.R.. Moscou, 1977 / W. RUDIN. *Real and Complex Analysis*. McGraw-Hill, New York, 3' éd. 1987 / E. WHITTAKER & G. N. WATSON, *A Course of Modern Analysis*, Cambridge Univ. Press, New York-Londres, 1969.

• Approximation des fonctions

C. M. BENDER & S. A. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, 1978 / P. DAVIS, *Interpolation and*

Approximation, rééd., Dover Publ., New York. 1975 / F. B. HILDEBRAND, *Introduction to Numerical Analysis*, repr. of 1974. *ibid.*, 1987 / G. G. LORENTZ. *Approximation of Functions*, 2e éd. 1985 / S. B. STECKIN, *The Approximation of Functions by Polynomials and Splines*, American Mathematical Society. Providence (R. I.), 1981 / M. ZAMANSKY, *Approximation des fonctions*, Hermann, Paris, 1985.

FONCTIONS ANALYTIQUES

Depuis l'Antiquité, on connaît en substance la série géométrique suivante :

$$\frac{1}{1-x} = 1 + x + x^2 + ... + x^n + ... (0 \leqslant x < 1).$$

Une des grandes découvertes qui jalonnèrent la formation du calcul infinitésimal au milieu du XVIIe siècle fut la possibilité de représenter les fonctions « usuelles » (logarithme, exponentielle, fonctions trigonométriques, etc.) par des développements en série analogues. Au XVIIIe siècle, la plupart des mathématiciens en étaient arrivés à ne plus guère considérer comme dignes d'intérêt que les fonctions dites « analytiques », égales à la somme d'une série convergente du type :

$$\sum_{n=0}^{\infty} c_n (x - x_0)^n$$

au voisinage de chacun des points x_0 de leur domaine d'existence.

La plupart des fonctions considérées au XVIIIe siècle étaient implicitement supposées analytiques, et J. L. Lagrange, dans sa *Théorie des fonctions analytiques* (1797), tente de fonder une théorie formelle de la dérivation, indépendante de la notion d'infiniment petit ou de limite, sur le concept de développement en série entière.

Si, depuis le début du XIXe siècle. il a fallu abandonner ce point de vue trop exclusif, et donner droit de cité à des fonctions bien moins « régulières » que les fonctions analytiques, le rôle de ces dernières reste fondamental dans toutes les parties des mathématiques. En fait. le concept d'analyticité s'est même élargi depuis le début du XXe siècle ; la définition d'une série entière :

$$\sum_{n=1} c_n x^n$$

et le calcul sur ces séries gardent en effet un sens lorsque les coefficients c_n et la variable x ne sont plus nécessairement des nombres réels ou complexes, mais plus généralement appartiennent à un *corps valué complet*, par exemple le corps des nombres *p*-adiques (cf. théorie des NOMBRES Nombres p-adiques) : il ne s'agit pas là d'une généralisation sans motivation. car ce qu'on appelle maintenant l'*analyse p-adique* a pris dans ces dernières années une importance de plus en plus grande dans toutes les questions touchant la théorie des nombres algébriques (cf. théorie des NOMBRES Nombres algébriques).

Toutefois, une étude plus poussée révèle que la possibilité de donner une définition de la notion de « fonction analytique » sur un corps valué complet quelconque K masque en réalité de profondes différences de comportement pour ces fonctions, selon la nature du « corps de base » K considéré. Déjà, entre le cas réel (K = R) et le cas complexe (K = C), il y a un clivage abrupt, la théorie des fonctions analytiques de variables complexes étant incomparablement plus riche et plus simple que celle des fonctions analytiques

de variables réelles (cf. représentation et approximation des FONCTIONS, chap. 1) : pour ne citer qu'un exemple, la seule existence de la dérivée première pour une fonction d'une variable complexe la rend ipso facto analytique. La raison de cette différence doit être cherchée dans la *théorie de Cauchy*, fondée sur le merveilleux outil que constitue l'intégrale curviligne des fonctions de variable complexe, d'une souplesse et d'une puissance incomparables.

C'est seulement de cette théorie et de ses extensions qu'il est question dans les articles qui suivent. Jusqu'à la fin du XIXe siècle, on ne s'est guère occupé que de la théorie des fonctions d'une seule variable complexe, qui offrait aux recherches un champ aussi vaste que riche (d'ailleurs loin d'être épuisé même à l'époque actuelle) : en dehors de ses innombrables applications dans toute l'analyse, la théorie de Cauchy rejoignait la géométrie différentielle par ses liens avec la représentation conforme et les surfaces minimales, la théorie du potentiel (en mathématiques et en physique) par ses étroites relations avec les fonctions harmoniques de deux variables, et (ce qui est sans doute le plus inattendu) l'arithmétique supérieure avec la théorie analytique des nombres (séries de Dirichlet, méthode de Hardy-Littlewood).

Les conceptions les plus profondes qui se font jour pendant cette période sont celles de Riemann, qui parvint à maîtriser les difficultés que créaient depuis Bernoulli et Euler les prétendues « fonctions multiformes » par l'invention géniale des « surfaces de Riemann », premier exemple de ce que l'on appelle maintenant les « variétés holomorphes », et point de départ de la topologie moderne. La nécessité d'admettre comme domaines de définition des fonctions holomorphes des variétés plus générales que les ouverts des espaces $\mathbf{C}^n$ devait apparaître avec encore plus de netteté dans les études sur les fonctions de plusieurs variables complexes. Cette étude ne commença guère qu'au début du XXe siècle ; dès les premiers travaux sur la question, on se rendit compte des différences profondes qui la distinguaient de la théorie des fonctions d'une seule variable complexe, qui apparaît maintenant comme un cas d'exception ; de ce fait, il fallut forger de nouveaux moyens d'attaque, qui ont surtout été l'œuvre de H. Cartan et K. Oka dans les années 1930-1955. La tournure essentiellement géométrique qu'a prise cette théorie lui vaut le nom amplement justifié de *géométrie analytique* sous lequel on la désigne aujourd'hui, et la rapproche étroitement de la *géométrie algébrique* moderne : le parallélisme des énoncés dans ces deux théories est tout à fait remarquable. mais les méthodes de démonstration sont en général très différentes, et beaucoup plus délicates pour la géométrie analytique.

Les liens entre la théorie des fonctions analytiques et la géométrie algébrique remontent d'ailleurs beaucoup plus haut, avec la théorie des fonctions elliptiques et des intégrales abéliennes, qui constituèrent dans la première moitié du XIXe siècle un des triomphes de la théorie de Cauchy et furent au centre des recherches de Riemann. La théorie des fonctions elliptiques conduisit un peu plus tard à l'étude de la fonction modulaire, premier exemple des *fonctions automorphes* d'une variable complexe, brillamment développée par Felix Klein et surtout Henri Poincaré. Par là même, la théorie des fonctions analytiques entrait en contact avec la théorie des groupes ; les rapports entre ces théories sont devenus encore plus étroits à l'époque

moderne, lorsque Siegel, en généralisant aux fonctions de plusieurs variables complexes la notion de fonction automorphe, a placé la théorie de ces dernières dans ce qui semble son cadre naturel, la théorie des espaces *symétriques* d'Élie Cartan.

JEAN DIEUDONNÉ



A. Fonctions analytiques d'une variable complexe

On se propose, dans cette partie, d'exposer, avec des démonstrations quasiment complètes, les résultats les plus élémentaires de la théorie des fonctions analytiques d'une variable complexe : les deux derniers chapitres sont consacrés à quelques résultats sans démonstration. Historiquement, l'extension au cas complexe de nombreuses fonctions classiques a été réalisée par l'intermédiaire des développements en série ; les séries entières restent à la base de l'étude locale des fonctions analytiques. Avec l'introduction de l'intégrale curviligne, on peut aborder des problèmes globaux, comme la recherche des primitives, qui font apparaître des conditions de nature « géométrique » ou, plutôt, topologique, imposées aux ouverts du plan complexe ; les représentations intégrales de Cauchy sont à la base du calcul des résidus, qui a d'innombrables applications pratiques.

On a passé sous silence les résultats relatifs aux fonctions harmoniques de deux variables, qui ne sont autres que les parties réelles de fonctions anaiytiques, en renvoyant à l'article POTENTIEL ET FONCTIONS HARMONIQUES.

1. Séries entières

La définition et l'étude des fonctions analytiques reposent sur la notion de série entière, c'est-à-dire de série de la forme :

$$(1) \qquad \sum_{n=0}^{\infty} a_n (z - a)^n,$$

où a et les a_n sont des nombres complexes donnés ; on dit qu'une telle série (1) est une *série entière* de centre a et de coefficients a_n.

On dit que la série (1) *converge normalement* dans un ensemble K C C si la série des modules de ses termes est uniformément convergente pour $z \in$ K. Rappelons qu'il suffit pour cela qu'il existe une série numérique *convergente* de terme général α_n telle que $a_n(z - a)^n| \leqslant \alpha_n$ pour tout $n \in \mathbf{N}$ et $z \in$ K.

On désigne, dans ce qui suit, par D (a. r) et $\overline{\mathrm{D}}$ (a, r) les disques ouvert et fermé de centre a et de rayon r, c'est-à-dire les ensembles de nombres complexes z tels que 3 $a < r$ et $|z$ $a \leqslant r$ respectivement.

Convergence

Étudions l'ensemble des nombres complexes z pour lesquels la série (1) est convergente. Posant Z $= z - a$ pour simplifier, on se ramène, par une translation, à une série entière :

$$(2) \qquad \sum_{n=0}^{\infty} a_n Z^n$$

de centre 0.

Théorème 1. Soit R (éventuellement égal à 0 ou à $+$ a?) défini par la *formule d'Hadamard* :

$$(3) \qquad \frac{1}{\mathrm{R}} \lim_{n \to \infty} (\sup_{p \geqslant n} a_p \ ^{1/p}) ;$$

alors, pour tout $r < R$, la série (2) converge normalement dans le disque fermé $\bar{D}(0, r)$ (en particulier, cette série converge absolument pour $|z| < R$) et diverge pour $|z| > R$.

Ce nombre R est appelé le *rayon de convergence* de la série entière (2) et le disque ouvert correspondant D(O, R), qui est éventuellement vide ou égal au plan complexe tout entier, est appelé le disque de convergence de cette série. Remarquons que le théorème n'affirme rien pour $|z| = R$; toutes les circonstances peuvent se rencontrer : divergence en tout point de ce cercle, convergence avec ou sans convergence absolue en certains points, convergence partout (cf. *infra*, *Principe des zéros isolés)*.

Il faut indiquer la démonstration du théorème 1. Supposons d'abord $R > 0$ et soit $r < R$: choisissons ρ tel que $r < \rho < R$. D'après (3), on a :

$$1/\rho > |a_n|^{1/n},$$

soit $|a_n| \rho^n < 1$, pour n assez grand. Pour n assez grand et $|z| \leqslant r$, on a donc :

$$|a_n z^n| \leqslant |a_n|\, r^n = |a_n|\, \rho^n \left(\frac{r}{\rho}\right)^n < \left(\frac{r}{\rho}\right)^n ;$$

ainsi le terme général de (2) est majoré en module pour $z \in \bar{D}(0, r)$ par le terme général d'une série géométrique convergente, d'où la convergence normale. Réciproquement, soit $|z| > R$. D'après (3), il existe alors pour n une infinité de valeurs telles que $|a_n|^{1/n} > 1/|z|$, soit $|a_n z^n| > 1$; ainsi le terme général de la série (2) ne tend pas vers 0 et cette série diverge.

Pour trouver un rayon de convergence, on utilise rarement la formule (3), dont l'intérêt est surtout théorique, mais on étudie, suivant les valeurs z, la convergence absolue de la série (2) en utilisant les critères classiques (cf. SÉRIES ET PRODUITS INFINIS). Ainsi, la série de terme général $n!\, z^n$ a un rayon de convergence nul (critère de d'Alembert) ; la série :

$$\sum_{n=0}^{\infty} \frac{1}{n!} z^n,$$

dont la somme est la fonction exponentielle complexe (cf. EXPONENTIELLE ET LOGARITHME, chap. 4), converge absolument pour tout z et a donc un rayon de convergence infini. Les séries :

$$\sum_{n=0}^{\infty} z^n, \quad \sum_{n=0}^{\infty} \frac{z^n}{n}, \quad \sum_{n=0}^{\infty} \frac{z^n}{n^2}$$

ont chacune un rayon de convergence égal à 1, mais ont des comportements très différents pour $|z| = 1$: la première ne converge en aucun point de ce cercle puisque alors son terme général est de module égal à 1 ; la deuxième converge (mais pas absolument) en tout point de ce cercle différent de 1 et diverge pour $z = 1$; la troisième converge normalement pour $|z| = 1$.

Fonctions analytiques

Soit U un *ouvert* du plan complexe et $f : U \rightarrow C$ une fonction à valeurs complexes. On dit que f est *analytique ou holomorphe* dans U si elle est développable en série entière au voisinage de tout point de U, c'est-à-dire si, pour tout $a \in U$, il existe une série entière de centre a dont la somme est égale à $f(z)$ dans un disque de centre a :

$$(4) \quad f(z) = \sum a_n (z - a)^n, \qquad |z - a| < r.$$

Remarquons que les coefficients a_n et le nombre $r > 0$ dépendent du point a considéré. A priori le nombre r est inférieur ou égal au rayon de convergence de la série

qui figure dans (4). Remarquons que cette série est normalement convergente dans tout disque fermé de rayon assez petit (théorème 1), et, par suite, f est une fonction continue dans U.

Voici, en liaison avec ce qui précède, un important exemple de fonction analytique. *Théorème 2.* La somme d'une série entière est analytique dans son disque de convergence.

On se ramène par translation à une série entière de centre 0 :

$$f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

de rayon de convergence R ; soit z_0 un point du disque de convergence et posons $z = z_0 + (z \quad z_0)$. La formule du binôme de Newton donne :

$$z^n = (z_0 + (z - z_0))^n = \sum_{p=0}^{n} C_n^p z_0^{n-p} (z - z_0)^p,$$

d'où :

$$f(z) = \sum_{n=0}^{\infty} a_n \left(\sum_{p=0}^{n} C_n^p z_0^{n-p} (z - z_0)^p \right).$$

On montre alors que, pour :

$$|z - z_0| < R - |z_0|,$$

la série double ci-dessus est sommable ; on peut donc intervertir les deux signes Σ de sommation, d'où :

$$f(z) = \sum_{p=0}^{\infty} \left(\sum_{n=p}^{\infty} a_n C_n^p z_0^{n-p} \right) (z - z_0)^p$$

$$= \sum_{p=0}^{\infty} c_p (z - z_0)^p,$$

avec z-z,, < R − z_0 |, en posant :

$$c_p = \sum_{n=p}^{\infty} a_n C_n^p z_0^{n-p}.$$

La fonction f est donc somme d'une série entière de centre z_0 dans le disque ouvert D(z_0, R z_0 |) ; remarquons que cette série a donc un rayon de convergence R −| z_0 |, mais ce rayon de convergence peut être strictement plus grand. Le disque de convergence « déborde » alors du disque de convergence de la série initiale, et cela permet de *prolonger* la fonction f en une fonction analytique dans la réunion des deux disques (cf. chap. 7).

Principe des zéros isolés

Examinons maintenant le comportement d'une fonction analytique au voisinage d'un point où elle s'annule.

Soit f une fonction analytique dans un ouvert U et $a \in$ U un zéro de f La fonctionf'est développable en série entière au voisinage de a, c'est-a-dire que l'on a (4) dans un disque D(a, r) où $a_0 = f(a) = 0$. Si tous les coefficients de la série (4) ne sont pas nuls, ce qui aura lieu en particulier si f n'est pas identiquement nulle dans D(a, r), désignons par a_k le premier coefficient non nul ; on a alors :

$$(5) \qquad f(z) = \sum_{p=k}^{\infty} a_n (z - a)^n$$

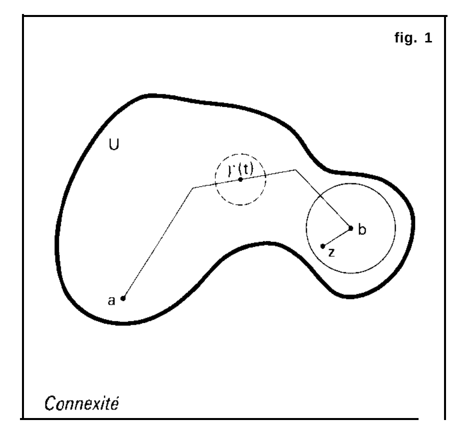
$$= (z - a)^k \sum_{p=0} a_{p+k} (z - a)^p$$

$$= (z - a)^k g(z),$$

où la fonction g est analytique dans D(a, r) et g(n) = $a_k \neq 0$. Le nombre entier ***k*** ainsi déterminé s'appelle, par analogie avec le cas des polynômes, l'*ordre du zéro* a. Puisque la fonction est continue et prend une valeur non nulle en a, il existe un disque D(a, p) C D(a, r) dans lequel elle ne s'annuie pas ; ainsi, $z = a$ est ie seui zéro de f dans le disque D(a, p). On a ainsi obtenu le *principe des zéros isolés,* que l'on

peut aussi énoncer ainsi : Soit f une fonction analytique dans un ouvert U ; si un point a $\in$ U est limite d'une suite de zéros de f, alors f est identiquement nulle dans tout un disque de centre a. Il en résulte aussi que, si la somme d'une série entière de centre a est identiquement nulle dans un disque de centre a, alors tous ses coefficients sont nuls ; on en déduit immédiatement l'*unicité* du développement en série entière d'une fonction analytique au voisinage de **tout** point.

Les résultats précédents sont dits locaux : ils s'appliquent à des disques de rayons assez petits. Pour « globaliser », nous aurons besoin d'introduire une notion topologique qui joue un rôle essentiel dans tout ce qui suit. On dit qu'un ouvert U est *connexe* s'il ne peut pas s'écrire comme une réunion de deux ouverts non vides disjoints ; un ouvert connexe est souvent appelé un *domaine.* Tout ouvert quelconque du plan est une réunion de domaines deux à deux disjoints appelés les *composantes connexes* de cet ouvert. La condition de connexité équivaut à ceci : deux points quelconques de U peuvent être joints par une ligne brisée entièrement contenue dans U (fig. 1) :



Connexité

supposons d'abord U connexe et soit $a \in$ U ; désignons par U_1 l'ensemble des points de U qui peuvent être joints à a par une ligne brisée située dans U et par U_2 l'ensemble des points pour lesquels il est impossible de trouver une telle ligne brisée. L'ensemble U_1 est ouvert, car, pour tout point $b \in U_1$, il existe un disque D de centre b contenu dans U, et ce disque est en fait contenu dans U_1, car tout $z \in$ D peut être joint à a par une ligne brisée dans U : il suffit de « rajouter » à la ligne brisée joignant a et b le segment d'origine b et d'extrémité z, qui est contenu dans D, donc dans U. Pour la même raison l'ensemble U_2 est aussi ouvert, car, si c $\in U_2$, il existe un disque de centre c contenu dans U, et ce disque est contenu dans U_2 ; en effet, s'il contenait un point $b \in U_1$, on pourrait joindre a à c par une ligne brisée en rajoutant le segment d'origine b et d'extrémité c à la ligne brisée joignant a à b, ce qui contredit CE U_2. Puisque U = $U_1 \cup U_2$, et U_1 non vide car il contient a. on a nécessairement $U_2 = \emptyset$. *Réciproquement,* faisons l'hypothèse que deux points de U peuvent toujours être joints par une ligne brisée située dans U ; on va supposer que U peut s'écrire $U_1 \cup U_2$, où U_1 et U_2 sont des ouverts non vides disjoints et aboutir à la contradiction qu'ils ne sont pas disjoints. Soit $a \in U_1$ et $b \in U_2$; il existe une ligne brisée L $\subset$ U image du segment [0, 1] par une fonction affine par morceaux $t \mapsto \gamma(t)$ d'origine a et d'extrémité b, c'est-à-dire telle que $\gamma(0) = a$ et $\gamma(1) = b$. Soit A l'ensemble des $t \in$ [O. 1] tels que $\gamma([0, t]) \subset$ U, et soit α la borne supérieure de A ; puisqu'il existe un disque ouvert de centre a contenu dans U_1, on a $\alpha > 0$. De plus, $\gamma(\alpha) \in U_1$. car, sinon, $\gamma(\alpha)$ appartiendrait à U_2 et, puisque U_2 est ouvert, il existerait tout Un intervalle $]\alpha - \eta, \alpha]$ dont l'image par γ serait conte-

nue dans U_2, en contradiction avec la définition de α ; il suffit donc de montrer que $\alpha = 1$, d'où $b = \gamma(1) \in U_1$, soit $U_1 \cap U_2 \neq \emptyset$. Or, si $\alpha < 1$, comme $\gamma(\alpha)$ appartient à U_1, qui est ouvert, il existerait tout un intervalle $[\alpha, \alpha + \eta'[$ dont l'image par γ serait contenue dans U_1, en contradiction avec le fait que α soit la borne supérieure de A.

Nous étant ainsi un peu familiarisés avec la notion d'ouvert connexe, on peut énoncer le *principe du prolongement analytique*. Soit f et g deux fonctions analytiques dans un ouvert *connexe* U ; s'il existe une suite de points distincts $z_n \in$ U convergeant vers $a \in$ U, avec $f(z_n) = g(z_n)$, alors on a identiquement $f(z) = g(z)$ dans tout U. Ce principe pourra s'appliquer par exemple si on a $f(z) = g(z)$ en tous les points d'une « ligne », ou, de manière encore plus particulière, si f et g sont égales dans un ouvert non vide inclus dans U. C'est la justification correcte du « prolongement des égalités » au domaine complexe : Si deux fonctions analytiques dans un même ouvert connexe U sont égales sur U $\cap$ R supposé non vide, alors elles sont égales dans U tout entier, ce qui permet d'étendre au champ complexe des formules connues dans le cas réel.

Établissons ce principe du prolongement analytique. Soit U_1 l'ensemble des points de U au voisinage desquels $f - g$ est identiquement nul, soit U_2 son complémentaire dans U. Tout point de U_1 est centre d'un disque dans lequel f-g est identiquement nul, et, par suite, tout ce disque est dans U_1 ainsi U_1 est ouvert et il est non vide, car d'après le principe des zéros isolés, il contient a. Montrons que U_2 est ouvert aussi, ce qui montrera, puisque U est connexe, que $U_2 = \emptyset$. Ceia résulte de la continuité en b, ou, si f(b) $\neq$ g (b), du principe des zéros isolés ; en effet, si $b \in U_2$, $f -$ g n'est pas identiquement nul au voisinage de b, et, par suite, il existe tout un disque de centre b dans lequel b est le seul zéro éventuel de $f -$ g, ce qui entraîne que ce disque est contenu dans U_2.

2. La dérivation complexe

Soit U un ouvert du plan et f une fonction à valeurs complexes définie dans U. On dit que f est *dérivable au sens complexe* en un point $z_0 = x_0 + iy_0 \in$ U si l'expression :

$$\frac{f(z_0 + u) - f(z_0)}{u}$$

tend vers une limite $f'(z_0)$ lorsque le nombre complexe $u = s + it$ tend vers zéro en module (c'est-à-dire lorsque (s, t) tend vers (0, 0) dans $\mathbf{R}^2$) ; le nombre complexe $f'(z_0)$ s'appelle la dérivée de f au sens complexe au point z_0.

Équations de Cauchy-Riemann

La condition de dérivabilité complexe au point z_0 peut aussi s'écrire :

$$f(z_0 + u) - f(z_0) = uf'(z_0) + u\,\varepsilon(u),$$

où $\varepsilon(u)$ tend vers 0 pour $|u| \to 0$. Si on pose $f(x + iy) = f(x, y)$, on aura :

$$f(x_0 + s, y_0 + t) - f(x_0, y_0) = (s + it)f'(z_0) + \varepsilon(s, t)\sqrt{s^2 + t^2},$$

ce qui exprime que la fonction $f(x, y)$, considérée comme fonction des deux variables réelles x et y, est dérivable (cf. CALCUL INFINITÉSIMAL - Calcul à plusieurs variables, chap. 2) et que :

$$(6) \quad \frac{\partial f}{\partial x}(z_0) = f'(z_0), \quad \frac{\partial f}{\partial y}(z_0) = if'(z_0)\ ;$$

d'où :

$$(7) \qquad \frac{\partial f}{\partial x}(z_0) + i\frac{\partial f}{\partial y}(z_0) = 0.$$

Réciproquement, si f est une fonction dérivable des deux variables x et y satisfaisant à la condition (7), elle est dérivable au sens complexe. Si on pose $f(x, y) = P(x, y) + iQ(x, y)$, la condition (7) donne les deux relations. appelées conditions de Cauchy-Riemann :

$$(8) \quad \frac{\partial P}{\partial x}(z_0) = \frac{\partial Q}{\partial y}(z_0), \quad \frac{\partial P}{\partial y}(z_0) = -\frac{\partial Q}{\partial x}(z_0).$$

Si f est une fonction dérivable des deux variables réelles x, y, on pose souvent :

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right), \quad \frac{\partial f}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right),$$

ce qui est suggéré par l'expression de la différentielle :

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \bar{z}}d\bar{z},$$

avec $dz = dx + i\,dy$ et $d\bar{z} = dx - i\,dy$. La condition de Cauchy-Riemann peut alors s'écrire :

$$\frac{\partial f}{\partial \bar{z}} = 0,$$

et on a, si f est dérivable au sens complexe :

$$f'(z_0) = \frac{\partial f}{\partial z}(z_0).$$

Dans ce qui suit, on va montrer qu'une fonction f est analytique dans un ouvert U *si et seulement si* elle est continûment dérivable au sens complexe dans U (cela signifie que f' est une fonction continue dans U). D'après ce qui précède, cela revient à dire que $f(z)$ est analytique si et seulement si $f(x, y)$ est une fonction continûment dérivable des deux variables x, y qui satisfait, en tout point $z = x + iy$ de U, à l'une des conditions équivalentes (7) ou (8).

Dérivation des fonctions analytiques

Montrons tout d'abord que *toute fonction analytique est indéfiniment dérivable au sens complexe* et que sa dérivée est encore une fonction analytique. La dérivabilité et l'analyticité étant des propriétés locales (cela veut dire que si tout point U est centre d'un disque dans lequel ces propriétés sont vraies, alors elles sont vraies dans U), il suffit d'établir le résultat suivant :

Théorème 3. Soit :

$$(*) \qquad f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n,$$

la somme d'une série entière dans un disque D(a, r) ; alors la fonction f est dérivable dans D(a, r), et on a :

$$(**) \qquad f'(z) = \sum_{n=1}^{\infty} n a_n (z-a)^{n-1}.$$

Remarquons tout d'abord que, puisque :

$$\lim_{n\to\infty} \sqrt[n]{n} = 1,$$

la formule d'Hadamard (3) montre que les séries entières $(*)$ et $(**)$ ont le même rayon de convergence. Par translation, on se ramène à a = 0.

Soit $z_0 \in$ D(O, r) et choisissons ρ tel que $|z_0| < \rho < r$; désignons enfin par $g(z)$ la somme de la série $(**)$ pour $|z| < r$. On a, pour $z \neq z_0$:

$$\frac{f(z) - f(z_0)}{z - z_0} - g(z_0) = \sum_{n=1}^{\infty} a_n \left[\frac{z^n - z_0^n}{z - z_0} - n z_0^{n-1}\right] ;$$

l'expression entre crochets est nulle pour $n = 1$ et, pour $n \geqslant 2$, on peut majorer son module :

$$\left|(z - z_0)\sum_{k=1}^{n-1} k z_0^{k-1} z^{n-k-1}\right| \leqslant |z - z_0| \frac{n(n-1)}{2} \rho^{n-2},$$

pour $|z| < \rho$. Par suite :

$$\left|\frac{f(z)-f(z_0)}{z-z_0}-g(z_0)\right| \leqslant |z-z_0| \sum_{n=2}^{\infty} n^2 |a_n| \rho^{n-2};$$

cette dernière série est convergente, puisque $\rho < r$, et f est donc dérivable en z_0, de dérivée $f'(z_0) = g(z_0)$.

Puisque f' est analytique, on peut lui appliquer de nouveau le théorème 3. Par récurrence, on obtient que f est indéfiniment dérivable au sens complexe et que sa dérivée k-ième est :

$$f^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1) \times (n-k+1) a_n (z-a)^{n-k},$$

pour $|z-a| < r$. Pour $z = a$, on a :

$$f^{(k)}(a) = k!\,a_k;$$

ainsi les coefficients a_k du développement (*) s'expriment simplement en fonction des valeurs des dérivées de f au point a. Si f est une fonction analytique dans un ouvert U, son développement en série entière de centre a est :

$$(9) \qquad f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(z-a)^n,$$

dans un voisinage de tout point a $\in$ U ; c'est la *formule de Taylor* de f au point a. Cela redémontre. en particulier, l'unicité de ce développement en série entière de centre a.

Il résulte du théorème 3 que, si :

$$f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n$$

dans un disque $D(a, r)$, la somme F(z) de la série entière :

$$F(z) = \sum_{n=0}^{\infty} \frac{a_n}{n+1}(z-a)^{n+1}$$

est une fonction analytique dans $D(a, r)$ qui admet f pour dérivée au sens complexe ; on dira que c'est une *primitive* de f au sens complexe. On a ainsi établi que toute fonction analytique possède *localement* une primitive définie à une constante additive près, c'est-a-dire que tout point de l'ouvert U dans lequel f est analytique est centre d'un disque dans lequel f admet une primitive. Mais l'existence d'une primitive n'étant pas une notion locale, on ne peut rien obtenir de plus pour l'instant. Il sera nécessaire d'introduire une nouvelle notion, de nature topologique, la simple connexité (cf. *infra*, *L'homotopie*, chap. 4), pour aborder globalement le problème des primitives d'une fonction analytique dans un ouvert.

Analyticité des fonctions dérivables

On se propose maintenant de montrer que toute fonction continûment dérivable (au sens complexe) dans un disque de centre a est somme dans ce disque d'une série entière. Avec le théorème 3, cela établira l'équivalence complète entre l'analyticité et la continue dérivabilité au sens complexe dans un ouvert U. Par translation, on se ramène à un disque ouvert de centre 0.

Théorème 4. Soit f une fonction continûment dérivable (au sens complexe) pour $|z| < R$; pour $r < R$, posons :

$$(10) \qquad a_n = \frac{1}{2\pi r^n}\int_0^{2\pi} f(re^{it})e^{-int}\,dt.$$

Alors les nombres a_n sont indépendants de r et on a :

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \quad |z| < R$$

Soit $z \in D(O, R)$; choisissons r tel que $|z| < r < R$. La fonction :

$$g(\lambda) = \int_0^{2\pi} \frac{f[(1-\lambda)z + \lambda re^{it}] - f(z)}{re^{it} - z} re^{it}\, dt$$

est continue et dérivable pour $0 \leqslant \lambda \leqslant 1$ (puisque, pour z fixé, la fonction sous le signe d'intégration est une fonction continûment dérivable de (t, λ)), et sa dérivée s'obtient par dérivation par rapport à λ sous le signe d'intégration. Ainsi :

$$g'(\lambda) = \int_0^{2\pi} f'[(1-\lambda)z + \lambda re^{it}] re^{it}\, dt = 0,$$

car la fonction sous le signe d'intégration est la dérivée (en t) de la fonction :

$$F(t) = \frac{1}{i\lambda} f[(1-\lambda)z + \lambda re^{it}],$$

qui est périodique de période 2π, d'où $g'(\lambda) = F(2\pi) - F(0) = 0$. Ainsi, g est constante dans $[0, 1]$, donc nulle puisque $g(0) = 0$. Écrivant que $g(1) = 0$, on obtient, en sortant $f(z)$:

$$(11) \quad f(z) \int_0^{2\pi} \frac{re^{it}}{re^{it} - z}\, dt = \int_0^{2\pi} \frac{re^{it}}{re^{it} - z} f(re^{it})\, dt.$$

Remarquons maintenant que, puisque $r > |z|$, on a :

$$(12) \quad \frac{re^{it}}{re^{it} - z} = 1 + z/re^{it} + (z/re^{it})^2 + \ldots + (z/re^{it})^n + \ldots,$$

où cette série est normalement convergente pour t réel. On peut donc intégrer terme à terme, ce qui donne :

$$(13) \quad \int_0^{2\pi} \frac{re^{it}}{re^{it} - z}\, dt = 2\pi,$$

puisque seul le terme constant a une intégrale non nulle. De même, $f(re^{it})$ est borné pour t réel ; on peut donc intégrer terme à terme son produit avec les deux membres de (12), ce qui donne, d'après (11), en tenant compte de (13) :

$$f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

les a_n ayant la valeur indiquée par (10). Le fait que les coefficients a_n sont indépendants de r résulte de l'unicité du développement en série entière.

Donnons une autre conséquence du théorème 4. Si f est analytique dans un ouvert U, pour **tout** point $a \in U$ elle est continûment dérivable dans le plus grand disque ouvert de centre a contenu dans U, donc développable en série entière de centre a dans ce disque ; par suite, le rayon de convergence de la série de Taylor de f en a est supérieur ou égal au rayon de ce disque, qui est la distance de a à la frontière de U (cf. chap 7).

Indiquons enfin qu'on peut établir le théorème 4 sous l'hypothèse plus faible que f est continue dans D(O, R) et dérivable au sens complexe en tout point, sans supposer la continuité de f' (théorème de Goursat) ; la difficulté est alors de montrer que la fonction $g(\lambda)$ qui figure dans la démonstration est encore dérivable, de dérivée nulle. Cela entraîne que, si f est dérivable au sens complexe en tout point d'un ouvert U, alors f est analytique dans U (donc indéfiniment dérivable au sens complexe).

3. Les coefficients de la série de Taylor

La formule (10) qui donne une expression intégrale des coefficients du développement en série entière va nous donner de précieux renseignements.

La propriété de moyenne

Considérons tout d'abord le terme constant de la formule de Taylor. On a, pour $n = 0$ dans (10) :

$$a_0 = f(0) = \frac{1}{2\pi}\int_0^{2\pi} f(re^{it})\,dt\,;$$

par translation, on aurait, pour tout point $a \in U$:

$$(14) \quad f(a) = \frac{1}{2\pi}\int_0^{2\pi} f(a + re^{it})\,dt,$$

pour tout r assez petit. Cette relation exprime que la valeur de f en a est égale à la valeur moyenne de f sur les cercles de centre a et de rayon r assez petit, ce qu'on exprime en disant que f possède la propriété de moyenne. Il est clair que la partie réelle et la partie imaginaire de f possèdent encore ces propriétés : ce sont des fonctions harmoniques (cf. POTENTIEL ET FONCTIONS HARMONIQUES).

Le principe du maximum

Une importante conséquence de la propriété de moyenne est le principe *du maximum*, que l'on peut énoncer ainsi : Soit U un ouvert *connexe* du plan complexe et f une fonction analytique dans U ; s'il existe un point $a \in U$ tel que l'on ait $|f(z)| \leqslant |f(a)|$ dans tout un disque de centre a (on dit alors que la fonction $|f|$ a un maximum relatif en a), alors f est constante dans l'ouvert U.

Pour cela, remarquons d'abord que, si f(a) = 0, alors on a $f(z) = 0$ dans tout un voisinage de a, d'ou $f \equiv 0$ dans U tout entier. Supposons donc $f(a) \neq 0$; multipliantfpar une constante complexe, on se ramène au cas où $f(a)$ est réel positif. D'après le principe du prolongement analytique, il suffit de montrer que f est constante dans un voisinage de a.

Soit R tel que $|f(z)| \leqslant f(a)$ pour $|z - a| < R$, et limitons-nous à des valeurs $r < R$; soit M(r) la borne supérieure de $|f(z)|$ pour $|z| = r$. D'après l'hypothèse, on a donc $M(r) \leqslant f(a)$. De plus, d'après la propriété de moyenne (14), on a $f(a) \leqslant M(r)$ et, par suite, $f(a) = M(r)$. Considérons la fonction $u(z)$ qui est la partie réelle de $f(a) - f(z)$, soit $u(z) = f(a) - \mathrm{Re} f(z)$; on a $u(z) \geqslant 0$, puisque $\mathrm{Re} f \leqslant |f|$ et $u(z) = 0$ si et seulement si $f(z) = f(a)$. Or u est la partie réelle d'une fonction holomorphe, donc possède la propriété de moyenne : sa moyenne sur le cercle $|z - a| = r$ est égale à $g(a)$, donc nulle :

$$\int_0^{2\pi} u(a + re^{it})\,dt = 0;$$

puisque u est continue $\geqslant 0$, cela entraîne $u(z) = 0$, donc $f(z) = f(a)$, dans tout le disque $D(a, R)$.

Voici par exemple une conséquence du principe du maximum qui sert souvent. Si f est une fonction analytique dans un disque D(u, R), alors on a, pour $r < R$:

$$(15) \qquad \sup_{|z-a|=r} |f(z)| = \sup_{|z-a|\leqslant r} |f(z)|\,;$$

en effet, la fonction continue $|f|$ atteint sa borne supérieure dans D(O, r) en un point qui, d'après le principe du maximum, est nécessairement un point frontière, sinon f serait constante (et alors (15) trivial). Ce raisonnement s'appliquerait à tout compact (ensemble fermé et borné) du plan : si f est analytique dans un ouvert connexe D d'adhérence $\overline{D}$ compacte et continue sur $\overline{D}$, alors $|f|$ n'atteint son maximum qu'en un point frontière, sinon elle est constante. Donnons une application de (15) en établissant le *lemme* ***de*** *Schwarz*, qui sert dans la théorie de la représentation conforme (cf. la partie C ci-après – Représentation

conforme) : Soit f une fonction analytique pour $|z| < 1$ telle que $f(0) = 0$ et $|f(z)| < 1$ pour $|z| < 1$; alors on a $|f(z)| \leqslant |z|$ pour $|z| < 1$, avec égalité en un point $z_0 \neq 0$, si et seulement si $f(z) = \lambda z$, λ constante complexe du module 1.

En effet, on a :

$$f(z) = \sum_{n=1}^{\infty} a_n z^n, \quad |z| < 1,$$

et, par suite, la fonction :

$$g(z) = \frac{f(z)}{z} = \sum_{n=1}^{\infty} a_n z^{n-1}$$

est analytique pour $|z| < 1$. Puisque $|f(z)| < 1$, on a donc $|g(z)| \leqslant 1/r$ pour $|z| = r$, et aussi pour $|z| \leqslant r$, d'après (15). Ainsi, fixant $z \in D(O, 1)$, on a $|f(z)| \leqslant |z|/r$ quel que soit $r > |z|$, $r < 1$; faisant tendre r vers 1, on a bien $|f(z)| \leqslant |z|$. Si maintenant on a $|f(z_0)| = |z_0|$ pour un point $z_0 \neq 0$, alors la fonction g atteint son maximum en un point du disque D(O, 1) ; la fonction g est donc constante et on a $f(z)/z = \lambda, |\lambda| = 1$.

Les inégalités de Cauchy

Soit f une fonction analytique dans un disque D(O, R) ; la fonction $f(z)$ est donc somme dans D(O, R) d'une série entière dont les coefficients a_n sont donnés par la formule (10). Si on désigne par M(r) le maximum de $|f(z)|$ pour $|z| = r$ (c'est aussi, d'après (15), le maximum pour $|z| \leqslant r$), on obtient donc :

$$(16) \qquad |a_n| \leqslant \frac{M(r)}{r^n}, \quad n \in \mathbf{N}.$$

Comme conséquence simple de (16), on obtient le *théorème de Liouville :* Une fonction analytique dans tout le plan et bornée est constante. En effet, cette fonction est alors développable en série de rayon de convergence infini ; les inégalités (16) entraînent que les coefficients a_n de ce développement vérifient :

$$|a_n| \leqslant \frac{M}{r^n},$$

quel que soit r. Pour r tendant vers l'infini, on obtient $a_n = 0$ pour $n \geqslant 1$. Le théorème de Liouville fournit une démonstration extrêmement simple du « théorème fondamental de l'algèbre », le *théorème de d'Alembert* : Tout polynôme à coefficients complexes de degré $\geqslant 1$ a au moins une racine complexe. En effet, soit :

$$P(z) = a_0 + a_1 z + \dots + a_{n-1} z^{n-1} + a_n z^n,$$

où $a_n \neq 0$, $n \geqslant 1$, un tel polynôme ; raisonnons par l'absurde, en supposant $P(z) \neq 0$ pour tout $z \in C$. La fonction $1/P(z)$ est alors analytique dans tout le plan et bornée, car $|P(z)| \to \infty$ pour $|z| \to \infty$; le théorème de Liouville entraîne alors que cette fonction est constante, en contradiction avec l'hypothèse.

4. Le problème des primitives

Nous avons obtenu maintenant à peu près tous les résultats qu'il est possible de démontrer *localement*, c'est-à-dire à partir de l'étude des séries entières ; pour continuer la théorie, nous avons besoin d'un remarquable outil introduit par Cauchy, l'intégrale curviligne le long d'une courbe.

l'intégrale curviligne

On va tout d'abord préciser la terminologie et les conditions de régularité auxquelles seront soumises les « courbes » du plan qui interviennent dans la suite.

On appelle *chemin* dans le plan complexe toute application continue $\gamma : I \rightarrow C$

d'un intervalle $I = [a, b]$ dans le plan complexe qui est continûment dérivable par morceaux ; cela signifie que I est une réunion d'un nombre fini d'intervalles fermés dans lesquels y est continûment dérivable, ou encore que γ est la primitive d'une fonction continue par morceaux dans I. Le point $\gamma(a)$ s'appelle l'*origine* du chemin y et le point $\gamma(b)$ est son *extrémité* ; un chemin ferme. c'est-à-dire tel que y(u) = y(b) sera appelé un *lacet*. Enfin, on appelle *trajectoire* le sous-ensemble y(I) du plan complexe parcouru par $\gamma(t)$ pour $t \in I$; si y est un chemin, on le représentera géométriquement en dessinant sa trajectoire et en indiquant par des flèches le « sens de parcours » du point y(t) lorsque t croît de a à b. L'exemple le plus simple d'une telle situation est celui d'une fonction affine par morceaux : la trajectoire est alors une ligne brisée (cf. chap. 1). Soit maintenant n un entier relatif non nul, z_0 un nombre complexe et r un nombre réel positif; pour $t \in [0, 2\pi]$, l'application :

$$t \mapsto z_0 + re^{2int}$$

est un lacet dont la trajectoire est le cercle de centre z_0 et de rayon Y (cf. EXPONENTIELLE ET LOGARITHME). Nous dirons que ce lacet est ce cercle « parcouru n fois », car tout point de ce cercle est l'image de n valeurs distinctes de $t \in [0, 2\pi]$; on dira aussi que c'est ce cercle parcouru une fois *dans le sens direct* si $n = 1$ et *dans le sens rétrograde* si $n = -1$.

Pour l'intégrale curviligne, nous aurons besoin d'une notion de courbe orientée plus « géométriquement », c'est-à-dire indépendante dans une certaine mesure du paramétrage y. Nous dirons que deux chemins :

$$\gamma_1 : I_1 \to \mathbf{C}, \quad \gamma_2 : I_2 \to \mathbf{C}$$

sont *équivalents* s'il existe une bijection continue φ croissante de 1, sur I_2, continûment dérivable par morceaux ainsi que la bijection réciproque, telle que $\gamma_1(t) = \gamma_2(\varphi(t))$ pour tout $t \in I_1$. Par exemple, pour tout chemin, on peut trouver, par une homothétie suivie d'une translation, un chemin équivalent défini dans un intervalle fixe de R, par exemple [0, 1].

Soit maintenant y : $[a, b] \to$ C un chemin et f une fonction à valeurs complexes définie et continue sur la trajectoire de y. D'après les conditions de régularité imposées à y, la fonction $t \mapsto f(\gamma(t))\gamma'(t)$ est continue par morceaux, donc intégrable, dans $[a, b]$. On appelle *intégrale curviligne* le long *du chemin* y le nombre complexe :

$$\int_\gamma f(z)\,dz = \int_a^b f(\gamma(t))\gamma'(t)\,dt\,; \tag{17}$$

par exemple, si y est le cercle de centre z_0 et de rayon r parcouru une fois dans le sens direct (cf. *supra)* et f une fonction continue sur ce cercle, on a :

$$\int_\gamma f(z)\,dz = \int_0^{2\pi} f(z_0 + re^{it})\,ire^{it}\,dt. \tag{18}$$

La notion d'équivalence des chemins introduite ci-dessus s'avère particulièrement bien adaptée à l'intégrale curviligne ; en effet, la formule de changement de variable dans une intégrale montre que, si y, et γ_2 sont des chemins équivalents et f une fonction continue sur leur trajectoire commune, les intégrales curvilignes de f sur y, et sur γ_2 sont égales.

Indiquons enfin comment on peut majorer une intégrale curviligne : si $|f(z)| \leqslant M$ pour tout point z de la trajectoire de y, on a :

$$\left|\int_\gamma f(z)\,dz\right| = \left|\int_a^b f(\gamma(t))\gamma'(t)\,dt\right| \leqslant M\int_a^b |\gamma'(t)|\,dt,$$

et cette dernière intégrale n'est autre que la *longueur* L(y) du chemin (cf. GÉOMÉTRIE DIFFÉRENTIELLE CLASSIQUE). Ainsi :

$$(19) \qquad \left|\int_a^b f(z)\,dz\right| \leqslant \mathrm{ML}(\gamma).$$

Lien avec les primitives

Si la fonctionfest définie dans un ouvert U contenant la trajectoire de y et est dans cet ouvert la dérivée au sens complexe d'une fonction continue F (d'après ce qui précède, F est alors nécessairement analytique), alors :

$$f(\gamma(t))\gamma'(t)$$

est, sauf pour un nombre fini de valeurs de t, la dérivée de la fonction continue $F(\gamma(t))$; par suite :

$$(20) \qquad \int_\gamma f(z)\,dz = F(\gamma(b)) - F(\gamma(a)).$$

En particulier, si une fonction f analytique dans un ouvert U admet une *primitive* **(au** sens complexe) dans cet ouvert, alors l'intégrale curviligne de f le long de tout lacet est nulle. Il est remarquable que, comme nous allons le montrer maintenant, cette condition est caractéristique : une fonction f analytique dans un ouvert *connexe* U admet une primitive dans U si et seulement si :

$$(21) \qquad \int_\gamma f(z)\,dz = 0$$

pour tout lacet y dont la trajectoire est contenue dans U.

Pour cela, remarquons tout d'abord que l'hypothèse (21) entraîne que, si y, et γ_2 sont deux chemins de U de *même origine* et *de même extrémité,* alors les intégrales curvilignes de f le long de y, et de γ_2 sont *égales.* En effet, si nous désignons par γ_2^0 le chemin « opposé » à γ_2, c'est-à-dire le chemin γ_2 « parcouru en sens inverse » (défini mathématiquement par :

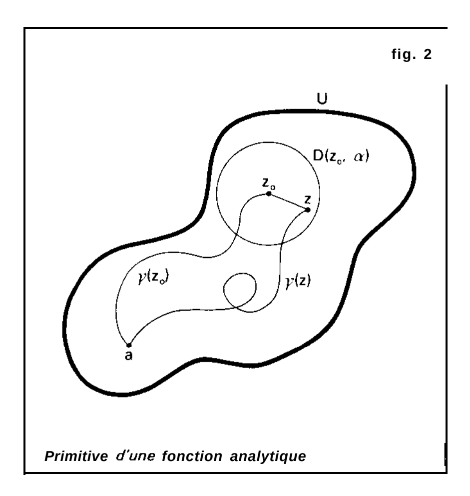
$$\gamma_2^0(t) = \gamma_2(c+d-t)$$

si $\gamma_2 : [c, d] \rightarrow$ C), l'intégrale de la fonction f le long de γ_2^0 est l'opposée de l'intégrale de f le long de γ_2. Le chemin obtenu par juxtaposition de y, et de γ_2^0, en parcourant « successivement » γ_1, puis γ_2^0, est alors un *lacet* ; d'où, d'après (21) :

$$0 = \int_{\gamma_1} f(z)\,dz + \int_{\gamma_2^0} f(z)\,dz$$
$$= \int_{\gamma_1} f(z)\,dz - \int_{\gamma_2} f(z)\,dz.$$

Pour f analytique dans U connexe satisfaisant à la condition (21), fixons un point $a \in$ U ; d'après ce qui précède, la valeur F(z) de l'intégrale curviligne de f le long d'un chemin y(z) de U d'origine a et d'extrémité z est indépendante du choix de ce chemin (il existe **au** moins un tel chemin qui soit une ligne brisée, cf. *supra*, chap. 1, *Principe* ***des*** *zéros isolés,* car U est connexe). Montrons que F est une fonction analytique qui a pour dérivée f dans U ; soit z_0 un point de U. 11 existe un disque $D(z_0, \alpha)$ dans lequel f est développable en série entière. Pour tout point z de ce disque, $F(z)$ est l'intégrale curviligne de f le long de tout chemin de U d'origine a et d'extrémité z, par exemple le chemin obtenu en parcourant successivement un chemin $\gamma(z_0)$ d'origine a et d'extrémité z_0 ; d'où une intégrale curviligne égale à $F(z_0)$, puis le segment $[z_0, z]$, soit :

$$F(z) = F(z_0) + \int_{[z_0, z]} f(\zeta)\,d\zeta$$



Primitive d'une fonction analytique

(fig. 2) ; or un calcul facile sur le développement en série entière defdans D(z,, α) montre que :

$$\int_{[z_0,z]} f(\zeta)\,d\zeta$$

est une primitive de f dans ce disque.

Remarquons que la condition (21) n'est pas toujours satisfaite. Par exemple la fonction $f(z) = 1/z$ est analytique dans l'ouvert connexe C* complémentaire de 0 dans le plan complexe, et son intégrale le long du cercle $|z| = 1$ parcouru une fois dans le sens direct est :

$$\int_0^{2\pi} e^{-it} ie^{it}\,dt = 2i\pi \neq 0;$$

ainsi, cette fonction n'admet pas de primitive dans C*.

On voit ainsi que la recherche d'une primitive d'une fonction analytique dans un ouvert U est un problème qui n'a pas toujours de solution. Pour aborder cette question nous aurons besoin d'introduire de nouvelles notions de topologie du plan. En particulier, on va voir que, si U satisfait à une condition de nature géométrique : être « simplement connexe » (cf. *infra, Théorème de Cauchy),* alors toute fonction analytique dans U a une primitive dans U.

L'homotopie

Soit U un ouvert du plan complexe et y, et γ_2 deux lacets de U ; quitte à remplacer l'un d'entre eux par un lacet équivalent, on peut supposer qu'ils sont définis dans le même intervalle I = [a, b] de R. On dit que ces deux lacets sont *homotopes dans* U s'il existe une application *continue :*

$$\varphi : I \times J - C,$$

J = [c, d] C R, telle que $\varphi(t, c) = \gamma_1(t)$, $\varphi(t, d) = \gamma_2(t)$ pour tout $t \in I$, et $\varphi(a, u) = \varphi(b, u)$ pour tout $u \in J$; intuitivement, cela signifie qu'on peut « passer » de y, à γ_2, sans sortir de U, par une « famille continue » de courbes *fermées* C,, (de représentation paramétrique $t \mapsto \varphi(t, u)$).

Cette notion d'homotopie des lacets est fondamentale dans toute la théorie de Cauchy ; on définit ainsi une relation d'équivalence sur l'ensemble des classes de lacets équivalents (au sens ci-dessus).

L'homotopie va permettre d'introduire une importante notion géométrique. On dit qu'un ouvert U du plan est *simplement* connexe si tout lacet de U est homotope dans U à un lacet constant, de trajectoire réduite à un point. Un exemple important est constitué par les ouverts *étoilés pur rapport à un point* ; cela signifie qu'il existe un point a d'un tel ouvert U tel que le segment $[a, z]$ soit entièrement contenu dans U pour tout $z \in$ U. C'est ainsi le cas du complémentaire dans le plan complexe d'une demi-droite fermée, si on prend pour a un point quelconque de la demi-droite opposée. Avec les notations précédentes, si U est un ouvert étoilé par rapport à a et si $\gamma : I \rightarrow$ U est un lacet de U, l'application :

$$\varphi : I \times [0, 1] \rightarrow U,$$

définie par :

$$\varphi(t, u) = u\gamma(t) + (1 - u)a,$$

réalise l'homotopie de y sur le lacet constant réduit au point a, car $\varphi(t, 0) = \gamma(t)$, $\varphi(t, 1) = a$; on peut écrire :

$$\varphi(t, u) - a = u(\gamma(t) - a),$$

ce qui montre que $\varphi(t, u)$ est homothétique de $\gamma(t)$ dans l'homothétie de centre a et de rapport u.

Le théorème de Cauchy

Ce résultat exprime que, si f est une fonction analytique dans un ouvert U, la valeur de l'intégrale curviligne defsur tout lacet de U ne dépend que de la classe d'homotopie de ce lacet dans U. En fait, on peut même montrer, ce qui est plus fort, qu'elle ne dépend que de la classe d'homologie (cf. *infra*, chap. 5, *L'indice*) de ce lacet. D'après ce qui précède, ce résultat est intimement lié au problème de la recherche de primitives « globales » pour une fonction analytique.

Théorème de Cauchy. Soit U un ouvert du plan complexe et f une fonction analytique dans U ; si y, et γ_2 sont des lacets de U qui sont homotopes dans U, alors :

$$\int_{\gamma_1} f(z)\, dz = \int_{\gamma_2} f(z)\, dz.$$

En particulier, si U est simplement connexe, on a :

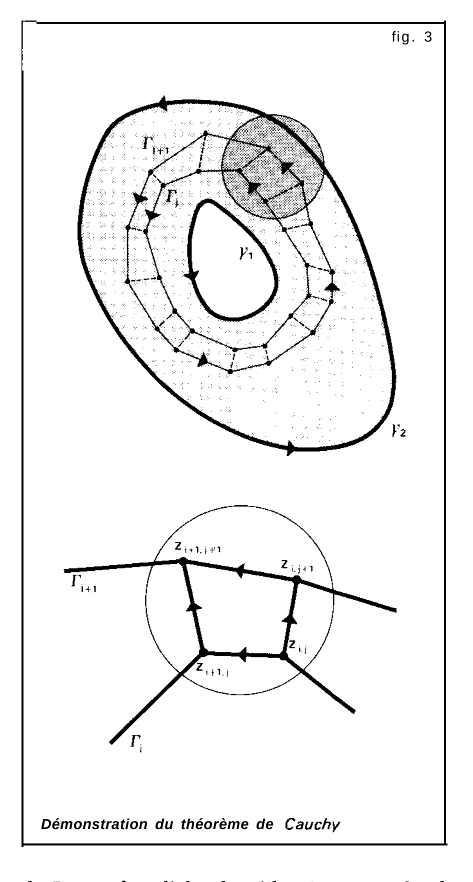
$$\int_{\gamma} f(z)\, dz = 0$$

pour tout lacet y de U ; par suite, toute fonction analytique dans un ouvert connexe et simplement connexe admet une primitive (au sens complexe) dans cet ouvert.

Indiquons, sans entrer dans tous les détails, les grandes lignes d'une démonstration de ce théorème. Soit, avec les notations ci-dessus :

$$\varphi : \mathbf{P} = [a,b] \times [c,d] \rightarrow \mathbf{U},$$

la fonction continue qui réalise l'homotopie. L'ensemble L = $\varphi(P)$ parcouru par $\varphi(t, u)$ pour (f, u) $\in$ P est compact et inclus dans U (fig. 3 ; L est ombré). La compacité



Démonstration du théorème de Cauchy

de L entraîne d'abord qu'il existe $\alpha > 0$ tel que, pour tout point $z_0 \in$ L, la fonction f soit développable en série entière en $(z - z_0)$ dans le disque D(z,, α) et, par suite (cf. *supra*, chap. 2, *Dérivation des fonctions*

analytiques), admette une primitive dans ce disque. D'après la continuité uniforme de φ sur P. il existe alors $\varepsilon > 0$ tel que $|t - t'| < \varepsilon$ et $|u - u'| < \varepsilon$ entraînent $|\varphi(t, u) - \varphi(t', u')| < \alpha/4$. Soit maintenant :

$$a = t_0 < t_1 < t_2 < ... < t_n = b,$$
$$c = u_0 < u_1 < u_2 < ... < u_m = d,$$

des subdivisions de $[a, b]$ et $[c, d]$ respectivement telles que $t_{i+1} - t_i < \varepsilon$ et $u_{j+1} - u_j < \varepsilon$. Posons $z_{i,j} = \varphi(t_i, u_j)$ et désignons par Γ_j la ligne brisée (fermée) joignant les points $z_{i,j}$ pour i croissant de 0 à n (fig. 3) : cette ligne brisée peut être paramétrée par une fonction affine par morceaux et nous désignerons encore par Γ_j le lacet ainsi défini.

Soit $A_{i,j}$ l'intégrale curviligne de f le long du segment $[z_{i,j}, z_{i+1,j}]$ et $B_{i,j}$ l'intégrale curviligne de f le long du segment $[z_{i,j}, z_{i,j+1}]$ (fig. 3). D'après le choix de ε et de la subdivision, les quatre segments correspondant aux intégrales $A_{i,j}$, $A_{i,j+1}$, $B_{i,j}$, $B_{i+1,j}$, sont dans un même disque où il existe une primitive F de f; on a donc :

$$\begin{aligned} A_{i,j+1} - A_{i,j} &= (F(z_{i+1,j+1}) - F(z_{i,j+1})) \\ &\quad - (F(z_{i+1,j}) - F(z_{i,j})) \\ &= (F(z_{i+1,j+1}) - F(z_{i+1,j})) \\ &\quad - (F(z_{i,j+1}) - F(z_{i,j})) \\ &= B_{i+1,j} - B_{i,j}. \end{aligned}$$

Additionnant les égalités ci-dessus pour $i = 0, 1, ..., n - 1$. on fait disparaître les $B_{i,j}$ et on obtient :

$$\int_{\Gamma_{j+1}} f(z)\, dz = \int_{\Gamma_j} f(z)\, dz,$$

puisqu'il est clair que :

$$\int_{\Gamma_j} f(z)\, dz = \sum_{i=0}^{n-1} A_{i,j};$$

un raisonnement très analogue montrerait que :

$$\int_{\Gamma_0} f(z)\, dz = \int_{\gamma_1} f(z)\, dz,$$
$$\int_{\Gamma_m} f(z)\, dz = \int_{\gamma_2} f(z)\, dz,$$

d'où le résultat cherché

Le logarithme complexe

Il résulte de ce qui précède que toute fonction analytique dans un domaine *simplement connexe* admet des primitives dans cet ouvert ; de plus, deux telles primitives diffèrent d'une constante au voisinage de chaque point (unicité du développement en série entière) et, par suite, dans l'ouvert connexe tout entier d'après le principe du prolongement analytique (cf. chap. 1, ***Principe des zéros isolés***). Ainsi, le choix de sa valeur en un point détermine entièrement cette primitive.

Voici une importante application de ce résultat. Si f est une fonction analytique dans un domaine simplement connexe U et ne *s'annule pas* dans U, la fonction f'/f est dérivable, donc analytique, dans U et, par suite, admet des primitives dans U. Soit L l'une d'entre elles, telle que l'on ait :

$$e^{L(a)} = f(a)$$

en un point $a \in U$; c'est toujours possible, car $f(a) \neq 0$ et tout nombre complexe non nul peut se mettre sous la forme e^b, $b \in \mathbf{C}$ (cf. EXPONENTIELLE ET LOGARITHME, chap. 4). La fonction analytique :

$$g(z) = e^{L(z)}$$

a pour dérivée :

$$g'(z) = L'(z) e^{L(z)} = \frac{f'(z)}{f(z)} g(z),$$

par définition de L. Ainsi :

$$f'(z)g(z) - f(z)g'(z) = 0;$$

divisant par $(f(z))^2 \neq 0$, on a :

$$\frac{f'(z)g(z) - f(z)g'(z)}{(f(z))^2} = 0,$$

ce qui exprime que la dérivée de la fonction f/g est nulle dans U, et, par suite, que cette fonction est constante dans le domaine simplement connexe U. Cette constante est égale à 1, puisque $f(a) = g(a)$, et on a donc :

$$g(z) = e^{L(z)} = f(z),$$

pour tout $z \in$ U. On dit que la fonction analytique L est une *détermination analytique* (en particulier continue) du logarithme de la fonction f.

Notamment, il existe des déterminations analytiques du logarithme de z dans tout domaine simplement connexe ne contenant pas 0 (par exemple, dans le complémentaire d'une demi-droite fermée issue de 0) ; et deux quelconques de ces déterminations diffèrent d'un multiple entier de $2\pi i$, mais on a vu plus haut qu'il n'existait pas de primitive de $1/z$ dans le plan complexe privé de 0 tout entier. Si L est une détermination du logarithme, sa partie imaginaire est une fonction continue qui est égale, en chaque point z, à une des valeurs possibles de l'argument de z. On dit que c'est une détermination continue de l'argument ; l'existence, dans un ouvert, d'une détermination continue $\arg z$ de l'argument équivaut à l'existence d'une détermination analytique du logarithme, par la formule :

(22) $$L(z) = \ln|z| + i \arg z,$$

où $\ln|z|$ est le logarithme du nombre réel positif $|z|$. On appelle *détermination principale du logarithme* de z la détermination analytique du logarithme, définie pour z non réel $\leqslant 0$, qui prend la valeur 0 pour $z = 1$; la détermination correspondante de l'argument prend ses valeurs entre $-\pi$ et $+\pi$ et la détermination principale réalise une bijection du plan complexe privé du demi-axe réel négatif sur la « bande » formée des nombres complexes $Z = X + iY$ tels que $\pi < Y < +\pi$ (cf. EXPONENTIELLE ET LOGARITHME, chap. 4). Pour z réel positif, L(z) est le logarithme de z au sens usuel, et la série de Taylor de cette fonction au point 1 est :

$$L(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} + \cdots + (-1)^{n+1}\frac{z^n}{n} + \cdots$$

Le théorème de Morera

Le théorème de Cauchy admet une réciproque, qui ne fait intervenir que des intégrales curvilignes le long de contours fermés très simples : le théorème de Morera affirme qu'une fonction f continue dans un ouvert U est analytique dans U si et seulement si son intégrale curviligne le long de tout rectangle, de côtés parallèles aux axes Ox et Oy, assez petit pour être entièrement contenu dans U est nulle. La condition nécessaire résulte du théorème de Cauchy ; indiquons le principe de la démonstration de la réciproque. Soit a un point quelconque de U et soit D un disque ouvert de centre a entièrement contenu dans U ; pour $z \in$ D, désignons respectivement par A, M, P, Q les points d'affixes respectives :

$$a = \alpha + i\beta, \quad z = x + iy,$$
$$x + i\beta, \quad \alpha + iy.$$

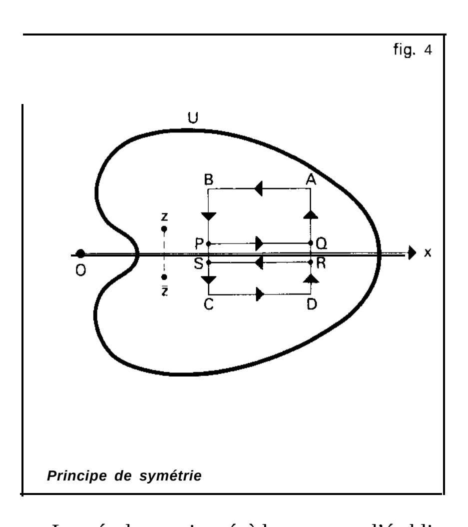
D'après l'hypothèse, les intégrales curvilignes de f le long des deux chemins APM et AQM sont égales ; soit $F(z)$ cette valeur

commune. On vérifie facilement que F admet dans D des dérivées partielles qui sont :

$$\frac{\partial F}{\partial x} = f, \quad \frac{\partial F}{\partial y} = if;$$

d'après les conditions de Cauchy-Riemann (6) ou (7), la fonction F est donc analytique, de dérivée au sens complexe f. Ainsi la fonction f, qui est, au voisinage de chaque point, la dérivée d'une fonction analytique, est analytique dans U.

Le théorème de Morera peut servir à éliminer des « fausses » singularités. Par exemple, désignons par P^+ et par P^- les demi-plans ouverts $\operatorname{Im} z > 0$ et $\operatorname{Im} z < 0$; si f est une fonction continue dans un ouvert U qui est analytique dans chacun des ouverts $U \cap P^+$ et $U \cap P^-$, alors f est en fait analytique dans U tout entier. Il suffit de montrer que l'intégrale de f le long de tout rectangle de côtés parallèles aux axes est nulle et le seul cas où une démonstration est nécessaire est celui où ce rectangle ABCD coupe l'axe réel, c'est-à-dire rencontre à la fois P^+ et P^- ; bornons-nous à de rapides indications géométriques. Introduisons, comme l'indique la figure 4, les segments opposés PQ et RS, parallèles et à une distance ε de l'axe réel. Puisque f est analytique dans P^+ et dans P^-, les intégrales curvilignes de f le long des lignes brisées QABP et SCDR sont respectivement égales aux intégrales curvilignes defle long des segments PQ et RS. La continuité de f entraîne alors que, lorsque ε tend vers 0, les intégrales de f le long de PS et de RQ tendent vers 0, tandis que les intégrales le long des segments PQ et RS tendent vers des valeurs opposées (fig. 4). A ia iimite, on obtient bien que l'intégrale de f le long du rectangle ABCD est nulle.



Principe de symétrie

Le résultat qui précède permet d'établir le *principe de symétrie* de Schwarz, utilisé pour prolonger des fonctions analytiques (cf. la partie B ci-après Fonctions elliptiques et modulaire). Soit U un ouvert connexe *symétrique par rapport à une droite* L ; désignons par V l'intersection de U avec un des demi-plans fermés déterminés par L et par U_1 l'intersection de U avec le demi-plan ouvert correspondant. Alors, toute fonction f continue dans V, analytique dans U, et telle que $f(U \cap L) \subset L'$, où L' est une droite, se prolonge (de manière unique car U est connexe) en une fonction analytique dans U tout entier (en faisant correspondre à des valeurs de z symétriques par rapport à L des valeurs de $f(z)$ symétriques par rapport à L'). En effet, par des transformations affines portant sur z et surf(s), on peut toujours se ramener au cas où L et L' sont l'axe réel, U symétrique par rapport à l'axe réel, $U_1 = U \cap P^+$ et $f(z)$ réel pour $z \in U \cap \mathbb{R}$. Il suffit alors de remarquer que la fonction g définie dans U tout entier par :

$$g(z) = f(z), \quad z \in U \cap P^+,$$
$$g(z) = \overline{f(\bar{z})}, \quad z \in U \cap P^-,$$

est continue dans U et analytique dans chacun des ouverts $U \cap P^+$ et $U \cap P^-$; remarquons que g prend des valeurs symétriques par rapport à l'axe réel en des points symétriques par rapport à l'axe réel.

5. La formule intégrale de Cauchy

Cette formule donne les valeurs d'une fonction analytique, sous forme d'une intégrale curviligne ; en particulier, elle traduit le fait que les valeurs d'une fonction analytique à l'« intérieur » d'une courbe sont déterminées par les valeurs prises sur la courbe. Nous aurons tout d'abord besoin de préciser une notion de caractère géométrique, le « nombre de fois » où une courbe fermée « tourne » autour d'un point.

L'indice

Soit $\gamma : I = [a, b] \to C$ un lacet et soit Ω le complémentaire de la trajectoire $\gamma(I)$ de γ. Pour tout $z \in \Omega$, on appelle *indice du point z* par rapport au lacet γ le nombre :

$$(23) \qquad j(z;\gamma) = \frac{1}{2i\pi}\int_\gamma \frac{d\zeta}{\zeta - z};$$

montrons que $j(z ; \gamma)$ est toujours un *entier relatif*. Pour $z \in \Omega$ fixé, posons :

$$h(t) = \int_a^t \frac{\gamma'(u)}{\gamma(u) - z}\,du \quad , \qquad t \in I,$$

de telle sorte que $j(z ; \gamma) = h(b)/2\,i\pi$, par définition de l'intégrale curviligne le long de γ. Puisque $e^w = 1$ si et seulement si $w/2\,i\pi$ est un entier (cf. EXPONENTIELLE ET LOGARITHME, chap. 4), il suffit d'établir que $e^{-h(b)} = 1$.

Posons :

$$g(t) = e^{-h(t)}(\gamma(t) - z);$$

puisque :

$$h'(t) = \frac{\gamma'(t)}{\gamma(t) - z},$$

sauf en un nombre fini de points, on a :

$$g'(t) = -h'(t)(\gamma(t) - z)e^{-h(t)} + \gamma'(t)e^{-h(t)} = 0,$$

ce qui montre que la fonction continue g est constante, sur [a, b]. En particulier, $g(a) = g(b)$, d'où, puisque $h(a) = 0$:

$$e^{-h(b)}(\gamma(b) - z) = \gamma(a) - z;$$

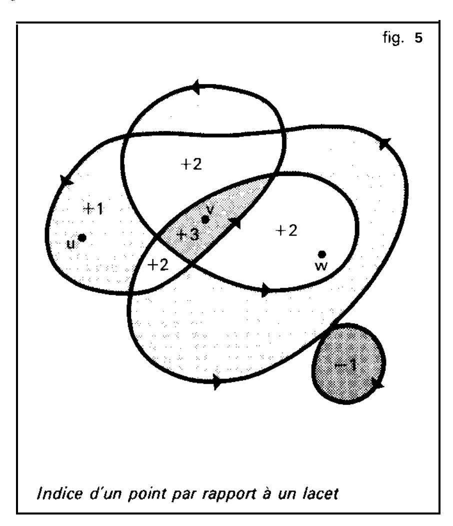
or $\gamma(b) = \gamma(a)$, car γ est un lacet, ce qui établit le résultat.

La formule (23) montre facilement que, pour γ fixé, la fonction $z \mapsto j(z ; \gamma)$ est continue dans Ω ; puisqu'elle ne prend que des valeurs entières, elle est donc constante dans chaque composante connexe de Ω. D'autre part, si γ_1 et γ_2 sont deux lacets homotopes dans le complémentaire de z, il résulte du théorème de Cauchy que $j(z ; \gamma_1) = j(z ; \gamma_2)$, puisque $\zeta \mapsto 1/(\zeta - z)$ est analytique dans $C - \{z\}$; en particulier, si la trajectoire de γ est contenue dans un ouvert simplement connexe U, on a $j(z ; \gamma) = 0$ pour tout $z \notin U$. Il en résulte que $j(z ; \gamma) = 0$ pour z assez grand et par suite dans toute la composante connexe non bornée de $\Omega = C - \gamma(I)$. Si $\gamma : t \mapsto e^{int}$, $n \in Z$, est le cercle unité parcouru n fois, on a :

$$j(0;\gamma) = \frac{1}{2i\pi}\int_\gamma \frac{d\zeta}{\zeta} = \frac{1}{2i\pi}\int_0^{2\pi} e^{-int}ine^{int}\,dt = n$$

et par suite $j(z;\gamma) = n$ pour $|z| < 1$, puisque le disque unité est connexe ; puisque l'extérieur du disque unité est connexe, on a $j(z ; \gamma) = 0$ pour $|z| > 1$. De manière générale, l'indice exprime le nombre algé-

brique de fois où le point $\gamma(t)$ « tourne » autour de z lorsque t croît de a à b. Par exemple, sur la figure 5, on a $j(u\,;\,$y$)=1$, $j(v\,;\,$y$)=3$, $j(w\,:\,$y$)=2$.



Indice d'un point par rapport à un lacet

La notion d'indice permet, en introduisant une nouvelle définition, de définir le cadre exact du théorème de Cauchy. Soit U un ouvert du plan ; on dira que deux lacets y et y' sont U-homologues si tout point du complémentaire de U a le même indice par rapport à ces deux lacets. Intuitivement, cela exprime que y et y' « tournent » le même nombre de fois autour de tout point du complémentaire de U. On voit alors facilement que deux lacets homotopes dans U sont a fortiori U-homologues. Le théorème de Cauchy, sous sa forme la plus générale, que nous admettrons, affirme que, si y et y' sont U-homologues, alors :

$$\int_{\gamma} f(z)\,dz = \int_{\gamma'} f(z)\,dz$$

pour toute fonction f analytique dans U.

Formule intégrale de Cauchy

Soit y un lacet dans un domaine *simplement connexe* U ; la formule intégrale de Cauchy exprime que, pour toute fonction f analytique dans U, on a :

$$(24)\qquad j(z\,;\gamma)f(z) = \frac{1}{2i\pi}\int_{\gamma}\frac{f(\zeta)}{\zeta - z}\,d\zeta,$$

pour tout point z de U n'appartenant pas à la trajectoire de y.

En effet, on voit facilement sur le développement en série entière de f au voisinage de z que la fonction g définie par :

$$\begin{cases} g(\zeta) = \dfrac{f(\zeta)-f(z)}{\zeta - z} \quad \text{pour } \zeta \neq z, \\ g(z) = f'(z) \end{cases}$$

est analytique dans un voisinage de z, et, par suite, dans U tout entier. D'après le théorème de Cauchy (cf. chap. 4), l'intégrale curviligne de g le long de y est donc nulle ; soit, puisque z n'appartient pas à la trajectoire de y :

$$\begin{aligned} 0 &= \int_{\gamma}\frac{f(\zeta)-f(z)}{\zeta - z}\,d\zeta \\ &= \int_{\gamma}\frac{f(\zeta)}{\zeta - z}\,d\zeta - f(z)\int_{\gamma}\frac{d\zeta}{\zeta - z} \\ &= \int_{\gamma}\frac{f(\zeta)}{\zeta - z}\,d\zeta - 2i\pi f(z)\,j(z\,;\gamma) \end{aligned}$$

La formule de Cauchy est particulièrement intéressante lorsque y est un cercle parcouru dans le sens direct ou rétrograde. car alors $j(z\,;\,$y$) = +1$ ou 1 pour tout point z intérieur à ce cercle.

La formule (24) exprime les valeurs d'une fonction analytique en fonction de sa restriction à la trajectoire d'un lacet y, qui est une fonction continue sur cette trajectoire. Réciproquement, on peut *définir* ainsi des fonctions analytiques. Plus précisément, soit y : $I \longrightarrow C$ un chemin (on ne suppose pas qu'il est fermé) et soit f une

fonction continue sur la trajectoire $\gamma(I)$; alors la fonction :

$$(25) \qquad g(z) = \int_\gamma \frac{f(u)}{u - z}\, du$$

est analytique dans l'ouvert complémentaire de $\gamma(I)$. La démonstration consiste à établir, par des majorations au voisinage de chaque point, que l'on peut dériver (au sens complexe) sous le signe d'intégration, ce qui entraîne l'analyticité ; ainsi, la dérivée n-iéme au sens complexe est donnée par :

$$(26) \qquad g^{(n)}(z) = n! \int_\gamma \frac{f(u)}{(u - z)^{n+1}}\, du,$$

obtenu en dérivant n fois par rapport à z. En particulier, si f est analytique dans un ouvert simplement connexe U, on a, en combinant les deux résultats précédents :

$$(27) \qquad j(z;\gamma) f^{(n)}(z) = \frac{n!}{2i\pi} \int_\gamma \frac{f(u)}{(u - z)^{n+1}}\, du,$$

pour tout lacet γ de U et tout $z \in$ U n'appartenant pas à la trajectoire de γ.

Suites convergentes de fonctions analytiques

Dans le domaine réel, une limite uniforme de fonctions dérivables n'est pas nécessairement dérivable ; en fait, le théorème de Weierstrass affirme même que toute fonction continue sur un intervalle borné $[a, b]$ est limite uniforme de polynômes (cf. représentation et approximation des FONCTIONS). Mais, dans le cas complexe, la situation est tout à fait différente, et la formule intégrale de Cauchy permet d'obtenir des conditions très fortes de régularité par passage à la limite.

La notion de convergence appropriée ici est la « convergence uniforme sur tout compact ». Soit $f_1, f_2, ..., f_n$, une suite de fonctions analytiques dans un ouvert U du plan complexe ; on dit que la suite (f_n) *converge uniformément sur tout compact vers une fonction* f si tout point $a \in$ U est centre d'un disque fermé Δ inclus dans U sur lequel la suite (f_n) converge uniformément vers f, c'est-à-dire :

$$\|f - f_n\|_\Delta = \sup_{z \in \Delta} |f(z) - f_n(z)|$$

tend vers zéro pour n tendant vers l'infini. On peut alors affirmer (théorème dû à Karl Weierstrass) que la fonction limite f est analytique dans U ; de plus, la suite (f'_n) des dérivées converge uniformément sur tout compact vers la dérivée f' de f. Soit en effet $\overline{D}(a, r)$ un disque fermé de centre a sur lequel la convergence est uniforme, et soit γ le cercle frontière $|z - a| = r$ parcouru une fois dans le sens direct. D'après la formule de Cauchy, on a :

$$f_n(z) = \frac{1}{2i\pi} \int_\gamma \frac{f_n(u)}{u - z}\, du,$$

pour $|z - a| < r$. Puisque les fonctions (f_n) convergent uniformément vers f sur le cercle $|z - a| = r$, un passage à la limite dans l'intégrale montre que :

$$f(z) = \frac{1}{2i\pi} \int_\gamma \frac{f(u)}{u - z}\, du,$$

ce qui établit, d'après (25), que f est analytique. Partant maintenant de :

$$f'_n(z) = \frac{1}{2i\pi} \int_\gamma \frac{f_n(u)}{(u - z)^2}\, du,$$

une majoration facile de $1/(u - z)^2$ montre alors que f'_n converge uniformément dans tout disque $|z - a| \leqslant r'$, $r' < r$, vers :

$$\frac{1}{2i\pi} \int_\gamma \frac{f(u)}{(u - z)^2}\, du = f'(z).$$

Les résultats précédents permettent de définir des fonctions analytiques à partir de séries ou de produits infinis de fonctions analytiques ou encore d'étendre (25) à des « chemins sans fin » (cf. fonction GAMMA).

6. Points singuliers isolés et résidus

On se propose ici, dans une première approche vers les points singuliers, d'étudier le comportement d'une fonction analytique dans un *disque pointé* $0 < |z - a| < r$, c'est-à-dire dans un disque ouvert privé de son centre ; si f ne se prolonge pas en une fonction analytique dans le disque entier, on dira que a est un *point singulier* (isolé) pour f.

La série de Laurent

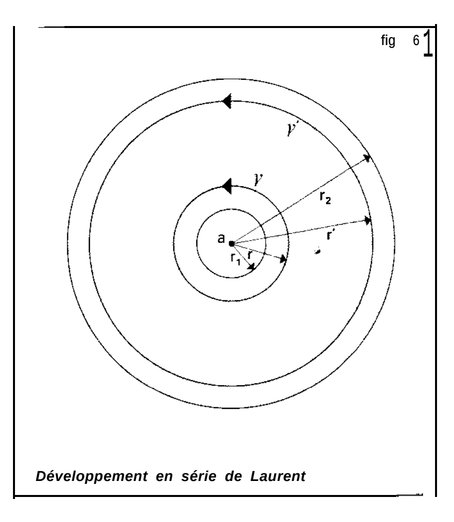
Un disque pointé $0 < |z - a| < r$ est un cas particulier d'une couronne ouverte $r_1 < |z - a| < r_2$; on va obtenir pour une fonction analytique dans une telle couronne S un développement en série généralisant le développement en série entière de centre a, valable seulement pour les fonctions analytiques dans tout un disque de centre a.

Il nous faut d'abord étendre la formule de Cauchy qui n'est pas applicable directement, puisque S n'est pas simplement connexe. Soit r et r' tels que $r_1 < r < r' < r_2$ et soit γ et γ' les cercles de centre a et de rayons r et r' parcourus une fois dans le sens direct (fig. 6). On voit comme ci-dessus que, pour $r < |z - a| < r'$, la fonction g définie par :

$$\begin{cases} g(u) = \dfrac{f(u) - f(z)}{u - z} & \text{pour } u \neq z, \\ g(z) = f'(z) \end{cases}$$

est analytique dans S. Puisque les lacets γ et γ' sont manifestement homotopes dans S, le théorème de Cauchy (cf. chap. 4) affirme que les intégrales de g le long de γ et γ' sont égales. On a donc :

$$\int_\gamma \frac{f(u)}{u - z} du - f(z) \int_\gamma \frac{du}{u - z} = \int_{\gamma'} \frac{f(u)}{u - z} du - f(z) \int_{\gamma'} \frac{du}{u - z},$$



Développement en série de Laurent

d'où, puisque $j(z ; \gamma) = 1$, $j(z ; \gamma') = 0$,

$$(28) \quad f(z) = \frac{1}{2i\pi} \int_{\gamma'} \frac{f(u)}{u - z} du - \frac{1}{2i\pi} \int_\gamma \frac{f(u)}{u - z} du,$$

qui généralise (24).

Remarquons alors que (en supposant, dans le calcul qui suit, $a = 0$ pour simplifier, ce qui ne retire aucune généralité), d'après le choix de z, $r < |z| < r'$, la série :

$$\frac{1}{u - z} = \frac{1}{u}\left[1 + \frac{z}{u} + \left(\frac{z}{u}\right)^2 + \ldots + \left(\frac{z}{u}\right)^n + \ldots\right]$$

converge uniformément pour $|u| = r'$, car alors $|u/z| = |z|/r' < 1$; on peut donc intégrer terme à terme sur γ' et on obtient :

$$(29) \quad f_1(z) = \frac{1}{2i\pi} \int_{\gamma'} \frac{f(u)}{u - z} du = \sum_{n=0}^{\infty} a_n z^n,$$

avec :

$$(30) \qquad a_n = \frac{1}{2i\pi} \int_{\gamma'} \frac{f(u)}{u^{n+1}} du,$$

en fait, ces coefficients a_n ne dépendent pas de r', car la fonction $f(u)/u^{n+1}$ est analytique dans S et, par suite, d'après le

théorème de Cauchy, on peut remplacer y' par n'importe quel cercle concentrique (contenu dans S) parcouru dans le sens direct, puisqu'il est homotope à y'. Ainsi la série entière définie dans (29) converge pour tout z de module $< r_2$; donc elle a un rayon de convergence $\geqslant r_2$ et, par suite, f_1 se prolonge en une fonction analytique pour $|z| < r_2$ (que nous désignerons encore par f_1). Remarquons qu'on aurait pu obtenir le résultat qui précède sans développer $1/(u-z)$ en série : d'après (25), la fonction f_1 définie par l'intégrale (29) est analytique pour $|z| < r'$; d'après le théorème 4 du chapitre 2 elle est donc développable en série entière et les intégrales (10) ne sont autres que des intégrales curvilignes du type (30) que l'on a explicitées en revenant à la définition. De même, la série :

$$\frac{1}{u-z} = -\frac{1}{z}\left(\frac{1}{1-u/z}\right) = -\frac{1}{z}\left[1+\frac{u}{z}+\frac{u^2}{z}+\dots+\frac{u^n}{z}+\dots\right]$$

converge uniformément pour $|u| = r < |z|$; on peut donc l'intégrer terme à terme sur y et on obtient :

$$f_2(z) = \frac{1}{2i\pi}\int_\gamma \frac{f(u)}{u-z}\,du = -\sum_{n=1}^{\infty} a_{-n}\left(\frac{1}{z}\right)n,$$

avec : $$a_{-n} = \frac{1}{2i\pi}\int_\gamma u^{n-1}f(u)\,du = \frac{1}{2i\pi}\int_\gamma \frac{f(u)}{u^{-n+1}}\,du\ ;$$

par le même raisonnement que ci-dessus, on voit alors que ces coefficients ne dépendent pas de r pour $r_1 < r < r_2$ et que la série de terme général $a_{-n}z^{-n}$ est convergente pour $|z| > r_1$ Sa somme, que nous désignerons encore par f_2, est donc une fonction analytique pour $|z| > r_1$.

Revenant au cas d'un point $a \in$ C, énonçons les résultats précédents. Soit S une couronne $r_1 < |z-a| < r_2$; toute fonctionfanalytique dans S peut s'écrire :

$$(31)\quad f(z) = \sum_{n=1}^{\infty}\frac{a_{-n}}{(z-a)^n} + \sum_{n=0}^{\infty} a_n(z-a)^n$$

où la première série converge pour $|z-a| >$ Y, et la seconde pour $|z-a| < r_2$; leurs sommes sont donc des fonctions analytiques pour $|z-a| > r_1$ et $|z-a| < r_2$ respectivement. Pour tout n entier relatif, les coefficients a_n sont donnés par :

$$(32)\qquad a_n = \frac{1}{2i\pi}\int_\gamma \frac{f(u)\,du}{(u-a)^{n+1}},\quad n \in z,$$

où y est un cercle quelconque de centre a et de rayon r, $r_1 < r < r_2$, parcouru une fois dans le sens direct.

Points singuliers

Soit f une fonction analytique dans un disque pointé $0 < |z-a| < r$. Un tel disque pointé étant un cas particulier de couronne, f admet un développement du type (3 1). Remarquons qu'ici la série entière :

$$u(\zeta) = \sum_{n=1}^{\infty} a_{-n}\zeta^n$$

converge pour tout $\zeta \in$ C ; u est donc une fonction analytique dans tout le plan telle que $u(0) = 0$. On dit que :

$$(33)\qquad u\left(\frac{1}{z-a}\right) = \sum_{n=1}^{\infty} a_{-n}\frac{1}{(z-a)^n}$$

est la *partie singulière* de la fonction f au point a. D'après (3 1), la fonctionfest donc somme de sa partie singulière et d'une fonction analytique dans tout le disque $|z-a| < r$.

Si $u = 0$, la formule (31) montre quef est somme d'une série entière pour $|z-a| < r$ et, par suite, se prolonge par continuité en une fonction analytique dans

tout ce disque : a n'est pas un « vrai » point singulier ; on dit que c'est un point *régulier.* Pour $u \neq 0$, on dit que a est un *point singulier isolé.*

Si u est un polynôme de degré $m \geqslant 1$, on dit que a est un *pôle* d'ordre m. Remarquons qu'alors la fonction $(z - a)^m f(z)$ se prolonge en une fonction h analytique dans le disque $|z - a| <$ r et que :

$$f(z) = \frac{h(z)}{(z-a)^m}, \quad h(a) \neq 0 ;$$

cela entraîne que $|f(z)| \to \infty$ pour $z \to a$. Plus généralement, si f et g sont analytiques dans un voisinage de a avec $f(a) \neq 0$ et a zéro d'ordre m de g (cf. chap. 1), alors f/g a un pôle d'ordre m en a, car on peut écrire g(2) $= (z - a)^m h(z)$, avec $h(z) \neq 0$, d'où f/h est analytique, dans un disque de centre a. Ce qui précède conduit à la notion de fonction méromorphe : Soit U un ouvert du plan et P un sous-ensemble fermé de U dont tous les points sont isolés ; on dit qu'une fonction analytique dans U – P est *méromorphe dans* U si les points de P sont des points réguliers ou des pôles (on peut d'ailleurs se ramener à ce seul cas en prolongeant F par continuité aux points réguliers ; P est alors l'ensemble des pôles de la fonction méromorphe). Remarquons que cela revient à dire que tout point de U est centre d'un disque pointé inclus dans U dans lequelf'est quotient de deux fonctions analytiques dans le disque tout entier.

Examinons enfin le cas où il existe une infinité d'entiers positifs n tels que $a_{-n} \neq 0$; on dit alors que a est un *point singulier essentiel.* Par exemple, la fonction :

$$e^{1/z} = \sum_{n=1}^{\infty} \frac{1}{n!z^n} + 1, \quad z \neq 0,$$

admet 0 pour point singulier essentiel. Un théorème dû à Émile Picard, précisant un théorème plus élémentaire de Weierstrass, décrit le comportement d'une fonction analytique autour d'un point singulier essentiel : Si a est un point singulier essentiel pour une fonction f, alors, dans tout disque pointé de centre a, la fonction f prend toutes les valeurs complexes, sauf au plus une ; par exemple, la fonction $e^{1/z}$ prend toutes les valeurs sauf 0 dans tout disque pointé de centre 0.

Le théorème des résidus

Soit U un ouvert et $a_1, a_2, \ldots, a_n$ des points distincts de U. Si f est une fonction analytique dans U' = U – $\{a_1, a_2, \ldots, a_n\}$, désignons par $v_1, v_2, \ldots, v_n$ les parties singulières de f en $a_1, a_2, \ldots, a_n$ respectivement ; ainsi, la fonction v_k est analytique dans le plan complexe privé du point a_k. La fonction :

$$g = f - (v_1 + v_2 + \ldots + v_n)$$

est alors analytique dans U' et sa partie singulière en chacun des points a_k est nulle ; elle se prolonge donc en une fonction, que nous désignerons encore par g, analytique dans U tout entier. Si U est simplement connexe, le théorème de Cauchy affirme que l'intégrale de g le long de tout lacet y est nulle, soit, si aucun des a_k n'appartient à la trajectoire de y :

$$(34) \qquad \int_\gamma f(z)\,dz = \sum_{k=1}^{n} \int_\gamma v_k(z)\,dz ;$$

ainsi, pour calculer l'intégrale de f, il suffit de calculer les intégrales des v_k. Effectuons ce calcul. Soit donc :

$$u(\zeta) = \sum_{n=1}^{\infty} a_{-n}\zeta^n$$

une fonction analytique dans tout le plan et y : I $\to$ C un lacet dont la trajec-

toire γ (I) ne contient pas un point a ; la série :

$$v(z) = \sum_{n=1}^{\infty} \alpha_{-n}(z-a)^{-n} = u\left(\frac{1}{z-a}\right)$$

converge uniformément pour $z = \gamma(t)$, $t \in I$, et, par suite, on peut intégrer terme à terme cette série sur le lacet γ. Or, pour $n \geqslant 2$, la fonction $(z-a)^{-n}$ admet une primitive $(1-n)(z-a)^{1-n}$ dans C — {a} et, par suite, son intégrale le long de γ est nulle ; il reste seulement le terme en $1/(z-a)$, qui introduit l'indice du point a par rapport au lacet γ :

$$\int_\gamma u\left(\frac{1}{z-a}\right)dz = 2i\pi\alpha_{-1}j(a\,;\gamma).$$

On voit donc l'importance du coefficient de $1/(z-a)$ dans la partie singulière au point a ; on l'appelle le *résidu* de f au point a, et on le note Res $(a\,;f)$. Ce qui précède permet d'énoncer le *théorème des résidus* sous sa forme générale : Soit U un ouvert simplement connexe et $a_1, ..., a_n$ des points distincts de U ; pour toute fonction f analytique dans :

$$U' = U - \{a_1, ..., a_n\}$$

et tout lacet γ de U', on a la *formule des résidus* :

$$(35) \quad \int_\gamma f(z)\,dz = 2i\pi \sum_{k=1}^{n} j(a_k\,;\gamma)\,\mathrm{Res}(a_k\,;f).$$

Cette formule est encore valable lorsque f est une fonction méromorphe dans l'ouvert simplement connexe U ; soit P l'ensemble des pôles de f. Comme P n'a que des points isolés dans U, pour tout lacet γ de U — P, l'ensemble des a ∈ P tels que $j(a\,;\gamma) \neq 0$ est fini. Dans ces conditions, on a :

$$(35') \quad \int_\gamma f(z)\,dz = 2i\pi \sum_{a\in P} j(a\,;\gamma)\,\mathrm{Res}(a\,;f).$$

Un exemple de calcul

La formule des résidus permet le calcul explicite de nombreuses intégrales. Donnons un exemple en calculant la transformée de Fourier de la fraction rationnelle $R(x) = 1/(x^2 + a^2)$, où a est un nombre complexe de partie réelle > 0. Il s'agit donc de calculer l'intégrale :

$$R(u) = \int_{-\infty}^{+\infty} R(x)e^{-2i\pi xu}\,dx, \quad u \in \mathbf{R}.$$

Considérons la fonction :

$$(36) \quad F(z) = \frac{e^{-2i\pi uz}}{z^2 + a^2} = \frac{e^{-2i\pi uz}}{(z-ia)(z+ia)},$$

qui est méromorphe dans tout le plan et admet ia et $-ia$ comme pôles d'ordre 1. Calculons les résidus en ces points.

De manière générale, si $F(z) = P(z)/Q(z)$, où P et Q sont analytiques dans un disque de centre b, $P(b) \neq 0$ et b *racine simple* de Q, on a vu que :

$$F(z) = \frac{\alpha_{-1}}{z-b} + h(z),$$

où α_{-1} est le résidu de f en b et h holomorphe au voisinage de b. Par suite, on a :

$$\mathrm{Res}(b\,;F) = \lim_{z\to b}(z-b)\frac{P(z)}{Q(z)} = \lim_{z\to b} P(z)\frac{z-b}{Q(z)-Q(b)} = \frac{P(b)}{Q'(b)},$$

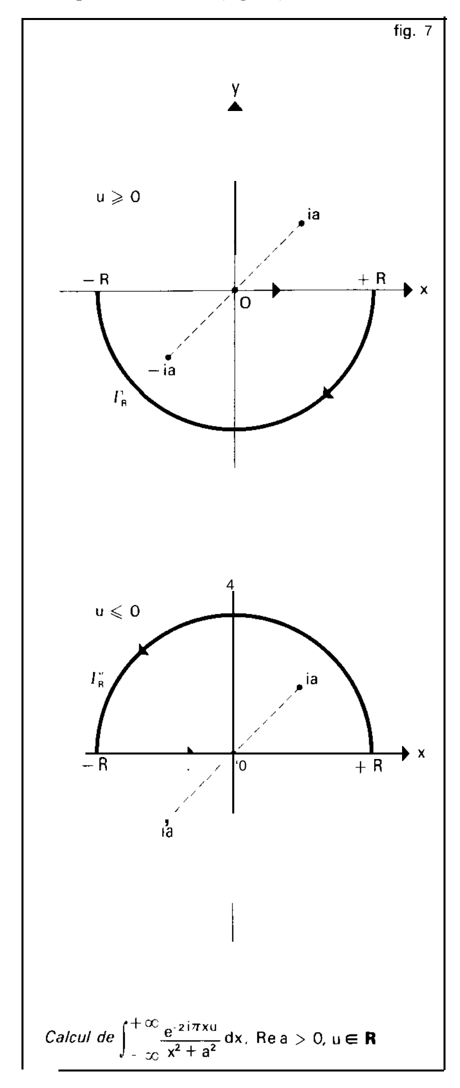
puisque $Q(b) = 0$. Dans notre cas, on a donc, F étant la fonction (36) :

$$\mathrm{Res}(ia\,;F) = \frac{e^{2\pi ua}}{2ia},$$

$$\mathrm{Res}(-ia\,;F) = -\frac{e^{-2\pi ua}}{2ia}.$$

Dans le calcul qui suit, on est alors amené à distinguer deux cas, suivant que $u \geqslant 0$ ou $u \leqslant 0$. Pour $u \geqslant 0$, appliquons la formule des résidus à l'intégrale de F le long du lacet constitué par le segment [- R, R] de l'axe réel et par le demi-cercle

Γ_R **de** centre 0 et de rayon R, R > $|a|$, du demi-plan inférieur (fig. 7). L'indice de ia



Calcul de $\int_{-\infty}^{+\infty} \frac{e^{-2i\pi xu}}{x^2+a^2}\,dx$, Re a > 0, u ∈ **R**

est égal à 0 et l'indice de ia est égal à 1. On a donc :

$$(37) \quad \int_{-R}^{R} \frac{e^{-2i\pi xu}}{x^2+a^2}\,dx + \int_{\Gamma_R} F(z)\,dz = -2i\pi \operatorname{Res}(-ia\,;F) = \frac{\pi}{a} e^{-2\pi au}\,;$$

or, sur Γ_R, on a :

$$|F(z)| \leqslant \frac{|e^{2\pi uy}|}{|z^2+a^2|} \leqslant \frac{1}{R^2-a^2},$$

car $|e^{x+iy}| = e^x$ et $y \leqslant 0$; la majoration (19) montre alors que l'intégrale de F le long de Γ_R tend vers 0 quand R tend vers l'infini, puisque :

$$\left|\int_{\Gamma_R} F(z)\,dz\right| \leqslant \frac{\pi R}{R^2-a^2},$$

d'ou, par passage à la limite dans (37) :

$$\int_{-\infty}^{\infty} \frac{e^{-2i\pi xu}}{x^2+a^2}\,dx = \frac{\pi}{a} e^{-2\pi au}, \quad u \geqslant 0.$$

Pour $u \leqslant 0$, on intègre sur le lacet constitué par le segment [-R, R] et le demi-cercle Γ'_R orienté comme l'indique la figure 7. L'indice de ia est cette fois égal à + 1, tandis que l'indice de $-ia$ est nul. On voit, comme ci-dessus, que l'intégrale curviligne de F le long de Γ'_R tend vers 0 pour R tendant vers l'infini et on obtient :

$$\hat{f}(u) = \int_{-\infty}^{\infty} F(x)\,dx = 2i\pi \operatorname{Res}(ia\,;F) = \frac{\pi}{a} e^{2\pi a|u|}, \quad u \leqslant 0.$$

En résumé, on a donc :

$$\hat{f}(u) = \frac{\pi}{a} e^{-2\pi a|u|}, \quad u \in \mathbf{R}.$$

Compteur logarithmique

Soit f une fonction méromorphe non nulle dans un ouvert simplement connexe U, soit N l'ensemble de ses zéros et P l'ensemble de ses pôles. Alors f'/f est une fonction méromorphe dans U et ses pôles, qui sont tous simples, sont les éléments de N ∪ P.

En outre, si $a \in$ N, on a :

$$\operatorname{Res}(a\,; f) = n_a,$$

où n_a est la multiplicité du zéro a, et, si $b \in$ P,

$$\mathrm{Res}(b\,;f) = -n_b,$$

où n_b est la multiplicité du pôle b.

La formule des résidus, appliquée à la fonction f'/f, montre que, pour tout lacet γ de U − (P ∪ N), on a :

$$\frac{1}{2i\pi}\int_\gamma \frac{f'(z)}{f(z)}\,dz = \sum_{a\in N} j(a\,;\gamma)n_a - \sum_{b\in P} j(b\,;\gamma)n_b.$$

C'est pourquoi cette fonction s'appelle compteur logarithmique des zéros et des pôles de f.

En particulier, si f est holomorphe dans U et si γ est le bord d'un disque contenu dans U − N :

$$\frac{1}{2i\pi}\int_\gamma \frac{f'(z)}{f(z)}\,dz = \sum_{a\in N_\gamma} n_a,$$

où N_γ est l'ensemble des zéros intérieurs au disque.

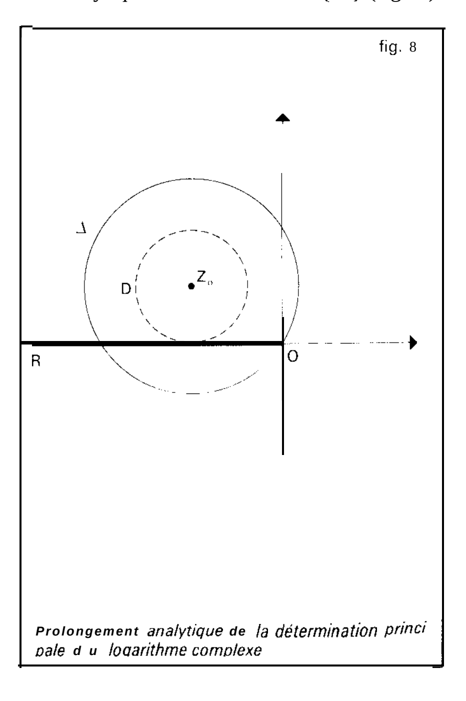
Cette formule permet en particulier de déterminer le nombre de zéros d'un polynôme, comptés avec leur ordre de multiplicité, contenus dans un disque donné.

Comme le compteur logarithmique dépend continûment des coefficients du polynôme, on en déduit que les racines d'un polynôme dépendent continûment des coefficients de ce polynôme (théorème de Weierstrass).

De manière semblable, on peut utiliser le compteur logarithmique pour établir des propriétés du spectre d'un endomorphisme d'un espace vectoriel de dimension finie sur C (continuité des racines du polynôme caractéristique).

7. Prolongement analytique

On se propose d'étudier ici la possibilité de prolonger une fonction f analytique dans un ouvert U à un ouvert plus grand. Pour tout point $a \in$ U, la série de Taylor de f en a converge dans le plus grand disque ouvert D contenu dans U (cf. chap. 2) ; mais, comme on l'a déjà signalé ci-dessus, il se peut fort bien que le disque de convergence Δ de cette série « déborde » de U. La somme de la série dans Δ est alors une fonction analytique dans Δ qui, d'après le principe du prolongement analytique (cf. chap. 1), coïncide avec f dans la composante connexe de U ∩ Δ qui contient D. Cependant, si U ∩ Δ n'est pas connexe, on ne peut pas affirmer en général que f se prolonge en une fonction analytique dans U ∪ Δ. Prenons pour exemple, pour U, le plan complexe privé des nombres réels négatifs ou nul U = C − $\mathbf{R}^-$ et, pour f, la détermination principale du logarithme (cf. chap. 4). Pour $z_0 = x_0 + iy_0$, $x_0 < 0$, $y_0 > 0$, le disque de convergence de la série de Taylor de $\ln z$ en z_0 est le disque de centre z_0 qui passe par 0, car la dérivée $1/z$ de $\ln z$ est analytique dans C* = C − {O} (fig. 8).



Prolongement analytique de la détermination principale du logarithme complexe

Ici $A \cap U$ a deux composantes connexes et la somme de la série de Taylor en z_0 n'est pas égale à $f(z)$ pour $z \in A \cap U$, Im $z < 0$, car Im $(\ln z) =$ Arg z subit une discontinuité en « traversant » $\mathbf{R}^- \cap A$, ce qui exclut la possibilité de prolonger f à $U \cup A$. L'examen de ces phénomènes a priori surprenants conduit à des extensions de la notion de fonction analytique et aux surfaces de Riemann. Pour éviter des difficultés du type précédent, nous raisonnerons ici sur des disques, car l'intersection de deux disques est toujours connexe.

Points singuliers et points réguliers

Soit f une fonction analytique dans un disque ouvert D ; on dira qu'un point frontière u est un *point régulier* pour f s'il existe un disque ouvert D_1 de centre u et une fonction g analytique dans D_1 qui coïncide avec f dans $D \cap D_1$. On peut alors prolonger f en une fonction analytique dans $D \cup D_1$, d'après le principe du prolongement analytique. Dans le cas contraire, on dit que u est un *point singulier pour* f. Il est clair que les points réguliers forment un ouvert de la frontière de D.

Remarquons que le fait, pour u, d'être un point régulier ou singulier n'a rien à voir avec la nature de la convergence de la série de Taylor de f en u. Par exemple,

$$f(z) = 1 + z + z^2 + \dots + z^n + \dots = \frac{1}{1-z},$$

pour $|z| < 1$, est analytique dans D(O, 1), et tous les points du cercle $|z| = 1$, sauf le point $z = 1$, sont des points réguliers, bien que la série diverge en **tout** point de ce cercle. On peut montrer que, si D est le disque de convergence d'une série entière de somme f, il existe *au moins* un point singulier pour f, sur sa frontière, mais il se peut que *tous* les points frontières soient des points singuliers (séries « lacunaires ») ; on dit alors que D est le domaine naturel d'existence de la fonction, car, dans ce cas, on ne peut pas la prolonger en une fonction analytique dans un ouvert connexe plus grand.

Éléments analytiques

Pour étudier les problèmes posés par le prolongement analytique, nous introduirons une nouvelle notion due, sous cette forme, à Weierstrass.

On appelle *élément de fonction analytique*, ou, en abrégé, *élément analytique* de centre a le couple (a, S) d'un nombre complexe a et d'une série entière S (de centre a) de rayon de convergence strictement positif; par abus de langage, on désignera encore par S, dans ce qui suit, la somme de la série S dans son disque de convergence. Par exemple, toute fonction analytique dans un ouvert U définit un élément analytique de centre a en tout point $a \in U$.

On dit que deux éléments analytiques (a, S) et (b, T) sont *le prolongement analytique direct* l'un de l'autre si leurs disques de convergence ont une intersection non vide et si S et T coïncident dans cette intersection. L'idée est de chercher à « prolonger » un élément analytique (a, S) en construisant des « chaînes » :

$$(a, S), (a_1, S_1), \dots, (a_n, S_n)$$

telles **que** deux éléments analytiques consécutifs quelconques soient prolongements directs l'un de l'autre.

Prolongement le long d'un chemin

Soit $\gamma : [0, 1] \rightarrow \mathbf{C}$ un chemin ; on dit qu'un élément analytique (b, T), de centre $b = \gamma(1)$, est le *prolongement analytique le long de* γ d'un élément (a, S) de centre $a = \gamma(0)$ s'il existe une chaîne (au sens précédent) :

$$(a, S), (a_1, S_1), \dots, (a_{n-1}, S_{n-1}), (a_n, S_n) = (b, T)$$

d'éléments analytiques centrés sur $\gamma([0, 1])$ dont les disques de convergence recouvrent cette trajectoire.

Si γ est un chemin d'origine a et (a, S) un élément analytique, il n'est pas toujours possible de prolonger cet élément analytique le long de γ (par exemple, si la trajectoire de γ contient un point singulier de S), mais on peut montrer que, si cela est possible, il y a *unicité* du prolongement. Par contre, ce prolongement ne dépend pas en général seulement de l'origine et de l'extrémité de γ. Considérons, par exemple, l'élément analytique $(1, S)$, où S est la série de Taylor au point 1 de la détermination principale du logarithme, $S(1) = 0$. Le prolongement analytique le long des deux demi-cercles $|z| = 1$, $\operatorname{Im} z > 0$ et $\operatorname{Im} z < 0$, respectivement de 1 vers -1, donne $(-1, T)$ et $(-1, U)$ où T et U sont des déterminations du logarithme dans $D(-1, 1)$ qui diffèrent de $2i\pi$, car :

$$T(-1) = i\pi$$

et :

$$U(-1) = -i\pi;$$

si on considère aussi, à partir de ce même élément $(1, S)$, les prolongements analytiques le long de tous les chemins d'origine 1 et d'extrémité -1 qui ne passent pas par 0, on obtient une infinité d'éléments analytiques différents $(-1, T_n)$, $n \in \mathbb{Z}$, où T_n est la détermination du logarithme dans $D(-1, 0)$ telle que :

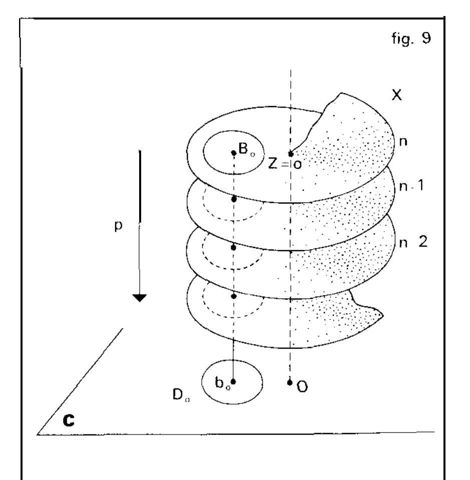
$$T_n(-1) = (2n + 1)i\pi.$$

Les problèmes que nous rencontrons ici et qui constituent ce qu'on appelait improprement la théorie des « fonctions » multiformes trouvent leur véritable cadre dans la théorie des surfaces de Riemann.

Surface de Riemann d'un élément analytique

Soit (a, S), un élément analytique de centre a, considérons l'ensemble X des éléments analytiques (b, T) qui sont des prolongements analytiques de (a, S) le long de chemins du plan complexe. On désignera par $p : X \rightarrow \mathbb{C}$ l'application de « projection » qui à $(b, T) \in X$ fait correspondre b.

Soit $B_0 = (b_0, T_0) \in X$ un élément analytique du disque de convergence D_0 et désignons par $V(B_0)$ l'ensemble des éléments analytiques centrés dans D_0 qui sont des prolongements analytiques *directs* de B_0 (ce sont les éléments analytiques déterminés par la fonction T_0 analytique dans D_0). La restriction p_0 de la projection p à $V(B_0)$ est une bijection de $V(B_0)$ sur D_0 au moyen de laquelle on peut « transporter » sur $V(B_0)$ la structure analytique du disque D_0. On *définira* ainsi, à partir de ceux de D_0, les ouverts de $V(B_0)$; une fonction à valeurs complexes définie dans $V(B_0)$ sera dite *analytique* si $f \circ p_0^{-1}$ est analytique dans D_0 (fig. 9) : par



Surface de Riemann du logarithme complexe au-dessus de $\mathbb{C}^* = \mathbb{C} - \{0\}$

exemple p est analytique ; de même, l'application qui à (b, T) associe $T(b)$ est analytique.

Par recollement, on a ainsi muni X d'une structure de variété analytique complexe de dimension (complexe) 1 ; le couple (X, ***p***) s'appelle la *surface de* ***Riemann*** de l'élément analytique (a, S). On dit aussi que c'est une surface de Riemann au-dessus de l'ouvert G = ***p(X)*** ; pour des exemples et une introduction à l'étude des surfaces de Riemann, nous renvoyons à l'article suivant, qui traite de la représentation conforme. Signalons cependant pour terminer ***le principe de monodromie,*** qui affirme que, si G est simplement connexe, alors p est une bijection, donc X est analytiquement isomorphe à G.

8. Théorèmes d'approximation

Le *théorème* **de** *Runge* affirme que, si U est un ouvert du plan complexe et A un ensemble ayant un point dans chaque composante connexe du complémentaire de U, alors toute fonction f analytique dans U est limite uniforme sur tout compact dans U d'une suite de fractions rationnelles dont les pôles appartiennent à A. Dans le cas particulier où U est borné et « sans trous », ce qui veut dire que C − U est connexe, alors f est limite uniforme sur tout compact d'une suite de polynômes.

Le *théorème* **de S.** *N. Mergelyan* est un résultat difficile. publié en 1954 ; il affirme que, si K est un compact du plan complexe, de complémentaire connexe, toute fonction continue sur K, holomorphe dans l'intérieur de K, est limite uniforme sur K de polynômes.

9. Théorèmes de décomposition

Théorème de factorisation de Weierstrass

La théorie de Weierstrass a pour objet de généraliser aux fonctions entières (c'est-à-dire analytiques dans tout le plan complexe) le théorème de d'Alembert-Gauss.

Les zéros non nuls d'une fonction entière peuvent être rangés en une suite (z_n), chaque zéro étant répété dans cette suite un nombre de fois égal à sa multiplicité, telle que $|z_n|$ soit une suite croissante, car tout disque ne contient qu'un nombre fini de tels zéros.

D'une part, étant donné une telle suite (z_n), il existe une fonction entière F associée à cette suite de zéros. L'idée la plus simple consiste à poser :

$$F(z) = \prod_{n=1}^{\infty} \left(1 - \frac{z}{z_n}\right),$$

mais ce produit ne converge que si la série $\sum 1/|z_n|$ est convergente. Dans le cas général, il faut « corriger » le facteur $(1 - z/z_n)$ par un facteur exponentiel. On est ainsi conduit à introduire les fonctions suivantes, dites facteurs primaires :

$$E_p(z) = (1 - z) \exp \left\{z + \frac{z^2}{2} + \dots + \frac{z^p}{p}\right\}.$$

Il est alors possible de trouver une suite (p_n) d'entiers positifs tels que le produit infini :

$$F(z) = \prod_{n=1}^{\infty} E_{p_n}\left(\frac{z}{z_n}\right)$$

soit normalement convergent sur tout compact de C, ce qui fournit une fonction qui convient.

Inversement, soit f une fonction entière, (z_n) la suite de ses zéros non nuls et k l'ordre de multiplicité de la racine 0. Alors,

la fonction $f/z^k F$ est une fonction entière qui ne s'annule pas, donc de la forme e^g, où g est entière.

Finalement, on a :

$$f(z) = z^k e^{g(z)} \prod_{n=1}^{\infty} E_{p_n}\left(\frac{z}{z_n}\right),$$

décomposition de f en facteurs primaires.

On peut étendre ce résultat à un ouvert U simplement connexe de C : Soit A un sous-ensemble discret (donc nécessairement dénombrable) de U et associons à chaque point a ∈ A une « multiplicité » n_a qui soit un entier positif; on peut alors montrer qu'il existe une fonction analytique dans U dont les zéros sont exactement les points de A, avec les multiplicités correspondantes. Il en résulte que, si h est une fonction *méromorphe* dans U, alors elle peut s'écrire, dans U tout entier et non pas seulement localement, comme quotient de deux fonctions analytiques.

Théorème de Mittag-Leffler

La théorie de Mittag-Leffler a pour objet de généraliser aux fonctions méromorphes dans C la décomposition en éléments simples des fractions rationnelles.

Soit fune fonction méromoruhe dans C et (z_n) la suite de ses pôles distincts, y compris éventuellement 0, organisée par module croissant, et soit :

$$P_n\left(\frac{1}{z - z_n}\right)$$

la partie principale relative au pôle z_n. Par un procédé correctif analogue à celui qui est employé dans le théorème de Weierstrass, on montre qu'il existe des polynômes Q_n, tels que la série :

$$F(z) = \sum_{n=1}^{\infty} P_n\left(\frac{1}{z - z_n}\right) - Q_n(z)$$

soit normalement convergente sur tout compact de C S, où S est l'ensemble des pôles.

Dans ces conditions, la fonction :

$$f(z) - F(z)$$

est une fonction entière.

On peut étendre ce résultat à un ouvert quelconque de C.

Les théorèmes de Weierstrass et de Mittag-Leffler s'appliquent notamment aux développements eulériens des fonctions transcendantes élémentaires (cf. EXPONENTIELLE ET LOGARITHME, chap. 5), au développement de la fonction gamma (cf. fonction GAMMA) et à la construction des fonctions elliptiques (cf. la partie B ci-après Fonctions elliptiques et modulaire).

JEAN-LUC VERLEY

Bibliographie

H. CARTAN, *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes*, Hermann, Paris, 6e éd. 1975 / B. V. CHABAT, *Introduction à l'analyse complexe*, vol. 1 : *Fonctions d'une variable*, Mir. Moscou, 1990 / J. DIEUDONNÉ, *Calcul infinitésimal, ibid.*, 2e éd. 1980 / P. DOLBEAULT, *Analyse complexe*, Masson. 1990 / M. HERVÉ, *Les Fonctions analytiques*, P.U.F., Paris, 1982 / E. HILLE, *Analytic Function Theory*, vol. 1. Ginn and Co.. Boston. 1959 / vol. II, Boston. 1962 / S. LANG, *Complexe Analysis*, Springer-Verlag, 2e éd. 1985 / R. NARASIMHAN, *Complex Analysis in One Variable*, Birkhaüser, Boston. 1985 / W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York. 3e éd. 1987.

B. Fonctions elliptiques et modulaire

Inaugurée par N. H. Abel et C. Jacobi, la théorie des fonctions elliptiques a été un sujet de prédilection pour les analystes pendant **tout** le XIXe siècle. Appliquées par B. Riemann et K. Weierstrass à l'étude des courbes algébriques dans le plan projectif

complexe, ces fonctions sont à la base de la théorie plus générale des *fonctions algébriques,* du domaine de l'algèbre et de la géométrie algébrique. Généralisées par H. Poincaré, qui a étudié les fonctions « fuchsiennes », elles sont aussi à l'origine de la théorie des *fonctions automorphes.* Il s'agit là de deux branches très actives des mathématiques contemporaines. qui utilisent simultanément des techniques d'analyse et d'algèbre très élaborées.

Intégrales circulaires et elliptiques

Le calcul intégral classique montre qu'une intégrale de la forme :

$$\int \frac{dx}{\sqrt{P(x)}},$$

où $P(x)$ est un polynôme du 2e degré sans racine double, se calcule à l'aide de fonctions dites élémentaires, c'est-a-dire circulaires ou hyperboliques. Posons par exemple :

$$u = \int_0^x \frac{dt}{\sqrt{1-t^2}};$$

si x et t sont réels, ils doivent être compris entre ± 1, et l'on a $u =$ Arc sin x, dont la fonction inverse est $x = \sin u$; comme u reste compris entre $\pm \pi/2$, la période 2π de cette fonction inverse n'apparaît pas si l'on prend x et t réels,

Mais prenons-les complexes : si ω est l'ensemble des points du plan dont l'affixe est non réel ou réel strictement compris entre ± 1. la fonction :

$$x \mapsto \frac{1}{\sqrt{1-x^2}}$$

a une détermination holomorphe sur ω, vaiant i à i-origine, qui à son tour a une primitive $u(x)$ holomorphe sur ω et nulle à l'origine. Quand x varie dans ω le long de la partie $[1, +\infty[$ (resp. $]-\infty, -1]$) de la frontière, au-dessus ou au-dessous, u décrit la droite Re $u = \pi/2$ (resp. $-\pi/2$) au-dessus ou au-dessous de l'axe réel. De la *formule intégrale de Cauchy* (cf. la partie A ci-dessus Fonctions analytiques d'une variable complexe, chap. 5) résulte alors une correspondance conforme biunivoque entre x décrivant ω et u décrivant la bande δ définie par :

$$-\pi/2 < \text{Re}\, u < \pi/2.$$

Le *principe de symétrie de Schwarz* (cf. la partie A ci-dessus - Fonctions analytiques d'une variable complexe, chap. 4) permet de prolonger cette correspondance par symétrie par rapport aux frontières rectilignes de ω et δ : après ce prolongement, à deux valeurs de u symétriques par rapport à l'une des droites Re $u = \pm \pi/2$ correspondent deux valeurs de x symétriques par rapport à l'axe réel, donc à deux valeurs de u différant de 2π correspond la même valeur de x. Ainsi l'inversion de l'intégrale circulaire :

$$u = \int_0^x \frac{dt}{\sqrt{1-t^2}},$$

effectuée dans le champ complexe, donne une fonction de période 2π, qui, d'autre part, est évidemment solution de l'équation différentielle :

$$\left(\frac{dx}{du}\right)^2 = 1 - x^2$$

Ce raisonnement, dont le principe est de Carl Jacobi (1804-1 85 1). s'applique aussi à l'intégrale elliptique :

$$\int \frac{dx}{\sqrt{P(x)}},$$

où P est le degré 3 ou 4, sans racine double. Prenons par exemple :

$$\int_x^\infty \frac{dt}{\sqrt{(t-\alpha)(t-\beta)(t-\gamma)}}$$

Cette intégrale a une détermination holomorphe sur ω, positive sur la partie $]\alpha, +\infty[$ de la frontière. Cette détermination, à son tour, a une primitive $u(x)$ holomorphe sur ω et nulle à l'infini. Quand x varie dans ω le long de la frontière, passant successivement par $+\infty$, α, β, y. $-\infty$, u décrit le périmètre 0, a, b, c, 0 d'un rectangle, où a et ic sont réels < 0 ; comme dans le cas précédent, la correspondance conforme biunivoque, entre x décrivant ω et u décrivant l'intérieur δ de ce rectangle, se prolonge par symétrie par rapport aux frontières rectilignes de ω et δ. Après ce prolongement, x prend la même valeur en deux points u symétriques par rapport à l'un des sommets du rectangle, donc admet un groupe (additif) de périodes engendré par $\tau = 2$ a, $\tau' = 2\ ic$, dont le rapport est imaginaire pur.

Ainsi l'inversion de l'intégrale elliptique :

$$u = \int_x^\infty \frac{dt}{\sqrt{(t-\alpha)(t-\beta)(t-\gamma)}}$$

donne une fonction doublement périodique, qui d'autre part est évidemment solution de l'équation différentielle :

$$\left(\frac{dx}{du}\right)^2 = (x-a)(x-\beta)(x-y).$$

Propriétés générales des fonctions analytiques uniformes admettant un groupe de périodes donné G

Le cas intéressant est celui qu'on vient de rencontrer, où G est engendré par deux périodes τ, τ' dont le rapport n'est pas réel. Une fonction holomorphe ou méromorphc (c'est-a-dire quotient de **deux** fonctions holomorphes) sur le plan complexe C, admettant le groupe de périodes G, peut être restreinte à un *parallélogramme de périodes* de sommets u, $u+\tau$, $u+\tau'$, $u+\tau+\tau'$, sur lequel elle prend toutes ses valeurs. ou bien considérée comme une fonction holomorphe ou méromorphe sur la variété compacte connexe C/G.

D'après le *principe du maximum*, cette fonction ne peut être holomorphe sans être constante ; on appellera donc G-elliptique une fonction f méromorphe sur C admettant le groupe de périodes G, ou bien méromorphe sur C/G. Si f prend la valeur x aux points $w_1(x), \dots$ M.,(X) de C/G (points distincts sauf pour une nombre fini de valeurs de x), on montre que l'entier n ne dépend pas de x, c'est *l'ordre* de la fonction f; d'après la formule intégrale de Cauchy prise le long du périmètre d'un parallélogramme de périodes, les images réciproques dans C des points w;(s), déterminées chacune modulo G, ont une somme indépendante (modulo G) de x. Une application géométrique de cette propriété est donnée dans l'article COURBES ALGÉBRIQUES, chapitre 7. Il résulte d'autre part que l'ordre de f est au moins 2.

Soit maintenant g une autre fonction G-elliptique, prenant les valeurs $y_j(x)$ aux points $w_j(x)$; le développement de :

$$\prod_{j=1}^{n} [g - y_j(x)]$$

est un polynôme en g de degré n dont les coefficients sont des fonctions rationnelles 1. $r_1(x), \dots$ $r_{,,}(x)$; on a donc, entre les deux fonctions G-elliptiques quelconques./ *et g.* la *relation algébrique* :

$$P(f,g) \equiv g^n + r_1(f)g^{n-1} + \quad + r_n(f) = 0$$

En particulier, f et sa dérivée f' sont liées par une relation algébrique : toute fonction G-elliptique est solution d'une équation différentielle algébrique.

Soit encore h une fonction G-elliptique, prenant les valeurs $z_j(x)$ aux points $w_j(x)$; le développement de :

$$\sum_{j=1}^{n} z_j(x) \prod_{k \neq j} [g - y_k(x)]$$

est encore un polynôme en g, cette fois de degré < n, dont les coefficients sont des fonctions rationnelles $s_1(x)$, ... s,(x) ; on a donc, entre les trois fonctions G-elliptiques f, g, h, la relation algébrique :

$$h P'_g(f, g) = s_1(f) g^{n-1} + + s_n(f),$$

d'où l'on peut tirer h en fonction *rationnelle* de f et g pourvu que $P'_g(f, g) \neq 0$. Ainsi les fonctions G-elliptiques sont exactement les fonctions rationnelles de deux d'entre elles, f et g, choisies de manière que, pour un x convenable, une valeur $y_j(x)$ soit distincte de toutes les autres. Les raisonnements de cet alinéa et du précédent peuvent être faits sur une variété compacte quelconque.

les fonctions de Weierstross

L'ordre d'une fonction G-elliptique étant au moins 2. on en cherche une d'ordre 2 : la somme de la série de terme général $1/(u - \tau)^2$, pour τ EG, répondrait à la question si elle avait un sens ; une légère modification. dont le seul but est d'assurer la convergence nécessaire, donne la fonction de Karl Weierstrass (18 15-l 897), notée par lui d'un p gothique :

$$(1) \quad \mathfrak{p}(u) = \frac{1}{u^2} + \sum_{\tau \in G - \{0\}} \left[\frac{1}{(u - \tau)^2} - \frac{1}{\tau^2} \right].$$

C'est une fonction paire et G-elliptique d'ordre 2, car l'origine en est un pôle double. et le seul pôle modulo G ; au voisinage de l'origine, on a :

$$(2) \quad \mathfrak{p}(u) = 1/u^2 + (g_2/20)u^2 + (g_3/28)u_4 +$$

avec :

$$(3) \qquad g_2 = 60 \sum_{\tau \in G - \{0\}} \frac{1}{\tau^4}, \quad g_3 = 140 \sum_{\tau \in G - \{0\}} \frac{1}{\tau^6}.$$

Du développement (2), il résulte que :

$$4\mathfrak{p}^3 - g_2 \mathfrak{p} - g_3 - \mathfrak{p}'^2$$

est holomorphe et nulle à l'origine, d'où la formule fondamentale :

$$(4) \qquad \mathfrak{p}'^2 = 4\mathfrak{p}^3 - g_2 \mathfrak{p} - g_3,$$

prouvant que x = $\mathfrak{p}$ (u) est fonction inverse de l'intégrale elliptique :

$$u = \int \frac{dx}{\sqrt{4x^3 - g_2 x - g_3}}$$

considérée au chapitre 1, ou encore que $x =$ p(u), y = p'(u) est une représentation paramétrique de la cubique :

$$y^2 = 4x^3 - g_2 x - g_3,$$

qui, pour cette raison, est dite elliptique (cf. COURBES ALGÉBRIQUES, chap. 7) .

Aux deux points de C/G, où p prend une valeur donnée x, la fonction p ' prend des valeurs opposées ; par suite (cf. chap. 2), les fonctions G-elliptiques sont exactement les *fonctions rationnelles* de p et $\mathfrak{p}'$. De la *formule de Weierstrass :*

$$\mathfrak{p}(u) + \mathfrak{p}(v) + \mathfrak{p}(u + v) = \frac{1}{4} \left[\frac{\mathfrak{p}'(u) - \mathfrak{p}'(v)}{\mathfrak{p}(u) - \mathfrak{p}(v)} \right]^2$$

résultent, d'abord, une relation algébrique entre p(u), $\mathfrak{p}(v)$, $\mathfrak{p}(u + v)$, puis une relation algébrique entre $f(u)$, $f(v)$, $f(u + v)$ pour une fonction G-elliptique quelconque f. Le *problème de Weierstrass* est la recherche des fonctionsfanalytiques uniformes ayant cette propriété : ce sont les fonctions rationnelles de la variable ou de l'exponentielle, et les fonctions elliptiques.

Les autres fonctions de Weierstrass attachées à G, ζ et σ, respectivement

méromorphe et holomorphe partout, ne sont pas G-elliptiques, mais sont liées à $\wp$ par les formules :

$$(5) \qquad \zeta = \quad \wp, \quad \sigma'/\sigma = \zeta\,;$$

elles permettent d'exprimer toute fonction G-elliptique d'ordre n soit en combinaison linéaire de n fonctions dérivées de translatées de ζ, soit comme quotient du produit de n translatées de σ par le produit de n autres translatées de σ.

Les fonctions de Jacobi

Les développements en séries de $\wp$ et ζ, en produit infini de σ, convergent lentement : si l'on garde seulement les termes ou facteurs correspondant aux périodes τ de modules $\leqslant k$, l'erreur commise est de l'ordre de $1/k$; le calcul numérique exige une convergence plus rapide, au moins celle d'une série géométrique, qu'on obtient en formant d'autres fonctions.

Choisissons un couple de périodes τ, τ' engendrant le groupe G : le rapport τ'/τ n'étant pas réel, on peut supposer sa partie imaginaire > 0, et, de plus, aussi grande que l'on veut, pour que le développement (6) ci-dessous converge plus vite ; alors A = exp $(\pi i\tau'/\tau)$ est le module < 1, aussi petit que l'on veut. On note :

$$(6) \qquad \mathrm{H}(u) = i \sum (-\ 1)^k \lambda^{(k-1/2)^2} \times \exp\,[(2k\quad 1)\pi iu/\tau],$$

somme portant sur tous les entiers k : c'est une fonction holomorphe partout, impaire, vérifiant :

$$(7) \qquad \mathrm{H}(u+\tau) = -\mathrm{H}(u),$$
$$\mathrm{H}(u+\tau') = -(1/\lambda)\exp(-\,2\pi iu/\tau)\mathrm{H}(u)\,;$$

elle ne dépend que des rapports τ'/τ et u/τ.

Par suite, la fonction méromorphe paire :

$$u \mapsto \exp(-\pi iu/\tau)\,\frac{\mathrm{H}(u+\tau/2)}{\mathrm{H}(u+\tau'/2)}$$

admet les demi-périodes τ et τ', et la fonction méromorphe impaire :

$$u \mapsto \exp(-\pi iu/\tau)\,\frac{\mathrm{H}(u)}{\mathrm{H}(u+\tau'/2)}$$

admet la demi-période τ et la période τ' ; on obtient les fonctions 2 G-elliptiques (2 G-homothétique de G dans le rapport 2) cn u et sn u en les multipliant respectivement par des constantes telles que cn 0 = sn $\tau/2$ = 1. La notation de ces nouvelles fonctions rappelle celle des fonctions circulaires cos et sin en raison de certaines analogies : ainsi $\mathrm{cn}^2 u + \mathrm{sn}^2 u = 1$; comme ces fonctions ne dépendent que des rapports τ'/τ et u/τ on peut choisir τ en fonction de τ'/τ de manière que les développements de Maclaurin commencent par :

$$\mathrm{cn}\,u = 1 - \frac{u^2}{2!} + \ldots.$$
$$\mathrm{sn}\,u = u - (k^2+1)\frac{u^3}{3!} + \ldots,$$

où k^2 ne dépend que de τ'/τ ; sn u est alors solution de l'équation différentielle :

$$\left(\frac{dx}{du}\right)^2 = (1-x^2)(1-k^2x^2),$$

donc s'obtient aussi par inversion de l'*intégrale elliptique de Legendre* :

$$u = \int_0^x \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}.$$

La fonction modulaire

Les formules (3) associent au groupe G les nombres g_2 et g_3, appelés *invariants de G* ; on peut en effet les considérer comme fonctions d'un couple τ, τ' de périodes

engendrant G. et ces fonctions sont inchangées quand on remplace le couple τ, τ' par un autre couple engendrant G, donc par un couple $a\tau$ + hr', $c\tau + d\tau'$, où a, b, c, d sont des entiers tels que ad bc = 1.

En outre. le rapport g_2^3/g_3^2 est conservé par une homothétie sur G, donc est fonction du rapport $\zeta = \tau/\tau'$ des deux périodes engendrant G, fonction inchangée quand on effectue sur la variable ζ une *substitution modulaire* :

$$\zeta \mapsto \frac{a\zeta + b}{c\zeta + d}$$

La *fonction modulaire* J est celle qui à $\zeta = \tau/\tau'$ fait correspondre :

$$J(\zeta) = \frac{g_2^3}{g_2^3 - 27 g_3^2};$$

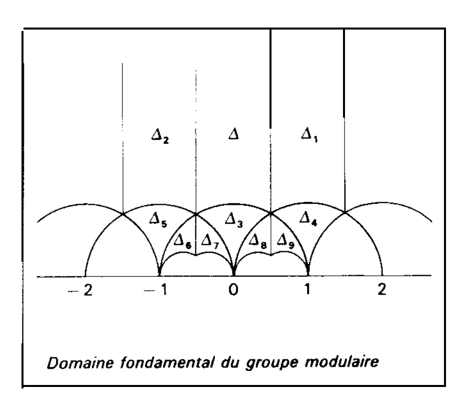
elle n'a de sens que pour ζ non réel, c'est pourquoi on la considère sur le demi-plan supérieur Im ζ < 0, où elle est holomorphe ; elle est invariante par les substitutions modulaires, en particulier par la translation $\zeta \mapsto \zeta + 1$. C'est donc aussi, pour $w|$ < 1, une fonction holomorphe de $w = \exp(2\pi i\zeta)$, à savoir :

$$(8) \quad J(\zeta) = \frac{1}{w \prod_{k \geqslant 1} (1 - w^k)^{24}} \times \left[\frac{1}{12} + 20 \sum_{k \geqslant 1} \frac{k^3 w^k}{1 - w^k}\right]^3.$$

Le *groupe modulaire*, formé des substitutions modulaires, opérant sur le demi-plan supérieur, admet le domaine fondamental A défini par les inégalités :

$$|\zeta| < 1, -1/2 < \mathrm{Re}\,\zeta < 1/2;$$

cela veut dire que les images de A par les substitutions modulaires (la figure ci-contre en indique queiques-unes) sont deux à deux disjointes tandis que les images de $\overline{\Delta}$ (réunion de A et sa frontière)



Domaine fondamental du groupe modulaire

couvrent le demi-plan. La jonction J réalise une bijection holomorphe de A sur l'ensemble des points du plan dont l'affixe est non réel ou réel > 1.

En particulier, la dérivée J' ne s'annule qu'aux points i etj de la frontière de A, où J prend les valeurs 1 et 0 respectivement, et aux points images des précédents par les substitutions modulaires ; par suite, la fonction analytique multiforme inverse de J, dont les valeurs appartiennent au demi-plan supérieur, se prolonge analytiquement le long de tout chemin, tracé dans le plan, évitant les points 0 et 1.

De là résulte que, pour les fonctions holomorphes omettant deux valeurs distinctes, donc aussi pour les fonctions méromorphes omettant trois valeurs distinctes, on retrouve certaines propriétés des fonctions holomorphes à valeurs dans un demi-plan ou, ce qui revient au même par composition avec une homographie, des fonctions holomorphes bornées. Ainsi, du fait qu'une fonction non rationnelle, méromorphe partout, ne peut être borner sur le complémentaire X d'un disque (théorème assez élémentaire, qui est dû à *Liouville*), on déduit, grâce à la fonction modulaire, qu'une teiie fonction ne peut omettre trois valeurs sur X (ce dernier théorème, beaucoup plus profond, est dû à *Picard*).

Les fonctions automorphes

On doit à Henri Poincaré (1854-1912) une vaste extension des fonctions elliptiques. Les translations étant des automorphismes du plan, c'est-à-dire des bijections holomorphes du plan sur lui-même, et les fonctions G-elliptiques des fonctions méromorphes sur le plan invariantes par le groupe G d'automorphismes, on peut de même se donner un groupe G d'automorphismes d'un disque ou demi-plan D et chercher des fonctions méromorphes sur D invariantes par G : on les appellera G-automorphes.

Pour qu'il en existe d'autres que les constantes, il est évidemment nécessaire que G satisfasse à la condition suivante, que Poincaré énonçait « G-discontinu », et qu'on énonce maintenant « G-discret » : aucun élément g de G n'est limite d'éléments de G distincts deg. Poincaré montra que cette condition nécessaire est aussi suffisante pour qu'il existe des fonctions G-automorphes (il disait « fuchsiennes ») non constantes.

Lorsque la variété D/G n'est pas compacte, ce qui est le cas général, deux fonctions G-automorphes ne sont pas en général liées par une relation algébrique : ainsi, pour le groupe modulaire, la fonction modulaire J est une fonction automorphe holomorphe, donc aussi e^J, qui n'est pas liée algébriquement à J. Une fonction automorphe pour ce groupe est liée algébriquement à J, si, et seulement si, comme J d'après la formule (8), cette fonction est une fonction méromorphe de $w = \exp(2\pi i\zeta)$ pour $|w| < 1$.

Poincaré caractérisa les domaines fondamentaux A des groupes G-discrets, et divisa ces groupes en familles suivant la disposition de A, dont dépend l'existence d'une relation algébrique entre deux fonctions G-automorphes. La première famille est formée des groupes G pour lesquels la frontière de A ne rencontre pas celle de D ; ce sont aussi les groupes G pour lesquels la variété D/G est compacte, de sorte que deux fonctions G-automorphes quelconques sont liées par une relation algébrique.

Inversement, à toute relation algébrique entre deux variables x et y, on peut associer un groupe G-discret et un couple de fonctions G-automorphes non constantesfet g liées par cette relation ; autrement dit, toute courbe algébrique peut être paramétrée à l'aide d'un couple de fonctions automorphes : $x = f(\zeta)$, $y = g(\zeta)$. Ce résultat remarquable de Poincaré, publié en 188 1, généralise le fait que toute cubique non unicursale, donc aussi toute courbe algébrique de genre 1 (cf. COURBES ALGÉBRIQUES, chap. 6 à 8), peut être paramétrée à l'aide d'un couple de fonctions elliptiques.

Les fonctions périodiques de plusieurs variables complexes

La construction de la fonction ℘ de Weierstrass montre, parmi bien d'autres choses, qu'étant donné deux nombres complexes τ_1, τ_2 linéairement indépendants sur le corps R des nombres réels il existe toujours une fonction méromorphe sur le plan C, dont le groupe de périodes est *exactement* celui qu'engendre le couple τ_1, τ_2.

11 n'en est plus ainsi lorsqu'on passe à m variables complexes, $m \geqslant 2$. Étant donné 2 m vecteurs $\tau_1, \dots, \tau_{2m}$ dans $\mathbf{C}^m$ (écrits dans la suite comme colonnes d'une matrice T à m lignes), linéairement indépendants sur R, pour qu'il existe une fonction méromorphe sur $\mathbf{C}^m$, dont le groupe de périodes soit exactement celui qu'engendrent $\tau_1, \dots, \tau_{2m}$, il faut et il suffit que ces 2 m vecteurs soient liés par les *conditions de Frobenius* suivantes : Il existe une matrice carrée A d'ordre 2 m, inversible et symétrique gauche, à éléments

entiers, telle que la matrice symétrique gauche (d'ordre m) TA(T) soit nulle, ce qui fait $m(m - 1)/2$ équations linéaires, et la matrice hermitienne (d'ordre m) i TA(T) définie positive.

D'autre part, le mathématicien Carl Ludwig Siegel a étendu à plusieurs variables le groupe modulaire, sous le nom de groupe modulaire symplectique. Au lieu d'une variable ζ telle que $(1/i)(\zeta - \overline{\zeta}) > 0$, il considère une matrice carrée symétrique Z d'ordre m, telle que la matrice hermitienne $(1/i)(Z - \overline{Z})$ soit définie positive : Z décrit alors, dans C^p avec $p = m(m + 1)/2$, un domaine de Siegel, qui est lié aux conditions de Frobenius, comme le demi-plan supérieur à la condition $\zeta = \tau_1/\tau_2$ non réel. De même que les automorphismes du demi-plan supérieur sont les homographies $\zeta \mapsto (a\zeta + b)(c\zeta + d)^{-1}$ avec a, b, c, d réels, $ad - bc = 1$, ou :

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix};$$

de même les automorphismes du domaine de Siegel sont les applications :

$$Z \mapsto (AZ + B)(CZ + D)^{-1},$$

avec A, B, C, D matrices carrées d'ordre m à éléments réels :

$$\begin{pmatrix} {}^tA & {}^tC \\ {}^tB & {}^tD \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

(1 est la matrice unité pour la multiplication) ; on obtient le groupe modulaire symplectique en prenant A, B, C, D à éléments entiers.

MICHEL HERVÉ

Bibliographie

P. APPEL & É. GOURSAT, *Théorie des fonctions algébriques et de leurs intégrales*, 3 vol., Gauthier-Villars, Paris, 1929-1930 / K. CHANDRASEKHARAN, *Elliptic Functions*, Springer-Verlag, New York, 1985 / C. HOUZEL, « Fonctions elliptiques et intégrales abéliennes », in J. Dieudonné et al., *Abrégé d'histoire des mathématiques*, t. II, chap. VII, Hermann, Paris / G. A. JONES & D. SINGERMAN, *Complex Fumions, an Algebraic and Geometric View Point*, Cambridge Univ. Press, 1988 / S. LANG, *Elliptic Functions*, Springer-Verlag, New York, 2ᵉ éd. 1987 / B. SCHOENEBERG, *Elliptic Modular Functions*, *ibid.*, 1974 / C. L. SIEGEL, *Topics in Complex Function Theory*, 3 vol., Wiley, New York, 1989 / G. VALIRON, *Cours d'analyse mathématique*, t. I : *Théorie des fonctions*, 3ᵉ éd., Masson, Paris, 1990 / A. WEIL, *Elliptic Functions According to Eisenstein and Kronecker*, Springer Verlag, New York-Berlin, 1976.

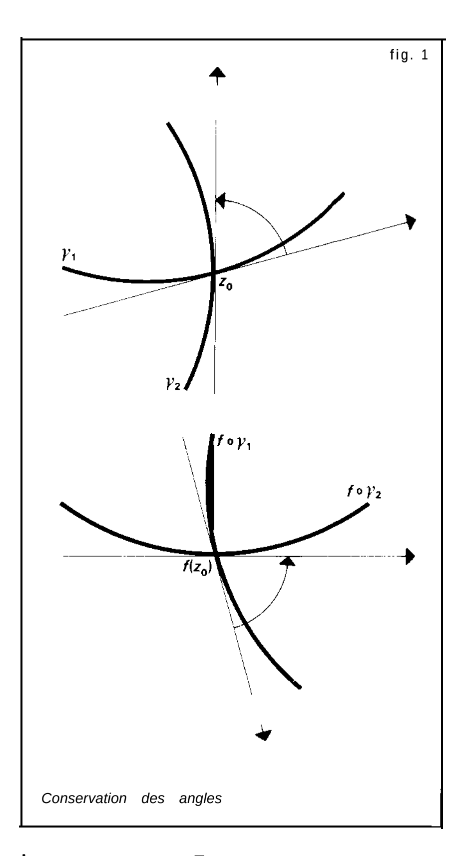
C. Représentation conforme

La représentation conforme la plus anciennement connue est la projection stéréographique, inventée par les Grecs (Hipparque, Ptolémée). Les problèmes cartographiques conduisirent à la découverte d'autres applications conservant les angles d'un domaine sphérique sur un domaine plan, telle la projection de Mercator (XVIᵉ siècle). Au début du XIXᵉ siècle, Carl Friedrich Gauss étudia systématiquement les propriétés intrinsèques des surfaces de l'espace habituel ; en particulier, il examina les applications bijectives d'une surface sur une autre qui sont différentiables, ainsi que leur réciproque, et qui conservent les angles. La notion de représentation conforme reçut un nouvel éclairage avec l'avènement de la théorie des fonctions d'une variable complexe, à laquelle elle est intimement liée. Bernhard Riemann sut exploiter cette relation de façon particulièrement féconde, introduisant la notion de *surface de Riemann*, qui résout les difficultés dues aux « fonctions multiformes » et donne un cadre convenable à la théorie du prolongement analytique. Cette théorie pose un certain nombre de problèmes topologiques qui ont conduit Bernhard Riemann et Henri Poincaré à développer les premières **bases** de la topologie algébrique.

1. Définition

La représentation conforme

Considérons un domaine D du plan $\mathbf{R}^2$. On dit qu'une application différentiable f de D dans $\mathbf{R}^2$ est conforme en un point z_0 de D si sa dérivée (ou application linéaire tangente) $D^1 f(z_0)$ en z_0 conserve les angles orientés (cf. CALCUL INFINITÉSIMAL - Calcul à plusieurs variables). En convenant que l'angle en z_0 de deux chemins différentiables γ_1 et γ_2 passant par z_0 est l'angle de leurs tangentes en z_0, on voit que cette condition revient à la suivante : l'angle orienté en $f(z_0)$ des chemins images $f \circ \gamma_1$ et $f \circ \gamma_2$ est égal à l'angle orienté de γ_1 et γ_2 en z_0, quels que soient les chemins γ_1 et γ_2 différentiables passant par z_0 (fig. 1).



Conservation des angles

On sait qu'une application linéaire du plan dans lui-même, qui conserve les angles orientés est une *similitude directe* de centre 0. Ainsi, la conformité de f en z_0 signifie que l'application linéaire tangente $D^1 f(z_0)$ est une similitude directe. Il est très commode de représenter les similitudes à l'aide de la multiplication des nombres complexes. Dans la suite, on considérera que le plan est le corps des nombres complexes C, et l'on écrira $x + iy$ pour le point (x, y) du plan (cf. nombres COMPLEXES) ; une similitude directe de centre 0 est alors une application de la forme $z \mapsto az$, où a est un nombre complexe non nul dont le module et l'argument sont respectivement le rapport et l'angle de la similitude ; dans la base canonique $(1, i)$ de C sur R, la matrice de la similitude considérée s'écrit :

$$\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

où α est la partie réelle de a, et β sa partie imaginaire.

Dire que f est conforme en z_0 revient donc à dire que sa dérivée est de la forme $h \mapsto ah$, avec $a \in \mathrm{C}$, $a \neq 0$; par conséquent, le rapport :

$$\frac{f(z_0+h)-f(z_0)-ah}{h} = \frac{f(z_0+h)-f(z_0)}{h} - a$$

tend vers 0 avec h, ou encore f est dérivable *au sens complexe* en z_0, avec comme dérivée, $f'(z_0) = a \neq 0$ (cf. la partie A ci-dessus -- Fonctions analytiques d'une variable complexe, chap. 2). En termes réels, on doit écrire que la matrice jacobienne de $f = P + iQ$, soit :

$$\begin{pmatrix} \dfrac{\partial P}{\partial x} & \dfrac{\partial Q}{\partial x} \\ \dfrac{\partial P}{\partial y} & \dfrac{\partial Q}{\partial y} \end{pmatrix}$$

est de la forme :

$$\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

ce qui donne les conditions de Cauchy-Riemann :

$$\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y}, \quad \frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x}.$$

Ainsi toute fonction holomorphefdans D, dont la dérivée ne s'annule pas, est conforme en tout point de D. Or on peut montrer que l'image d'une partie ouverte de C par une fonction holomorphe non constante est ouverte ; l'image du domaine D par une fonction holomorphe non constantefest donc un domaine $f(D)$. De plus, si f est injective (on dit quelquefois univalente), sa dérivée ne s'annule pas ; f définit une bijection de D sur $f(D)$ dont l'application réciproque est holomorphe dans $f(D)$; f est alors une *représentation conforme* de D sur f'(D). Les domaines D et f (D) sont dits conformément équivalents, ou encore isomorphes ; en ce qui concerne la théorie des fonctions analytiques, ils ont les mêmes propriétés, car l'application : $g \mapsto g \circ f$ est une bijection de l'ensemble des fonctions holomorphes dans f (D) sur l'ensemble des fonctions holomorphes dans D. Enfin, si g est une représentation conforme de $f(D)$ sur un nouveau domaine D', la composée $g \circ f$ est une représentation conforme de D sur D'.

Exemples de représentations conformes

Chaque fonction holomorphe injective dans un domaine D définit une représentation conforme de D sur f (D). Par exemple, la fonction $z \mapsto z^2$ est holomorphe et injective dans le *demi-plan supérieur*, défini par Im $z > 0$: **son** image est le complémentaire dans C de $\mathbf{R}_+^*$ (ensemble des nombres réels strictement positifs), c'est-à-dire le plan fendu suivant le demi-axe réel positif. Comme :

$$(x + iy)^2 = x^2 - y^2 + 2ixy,$$

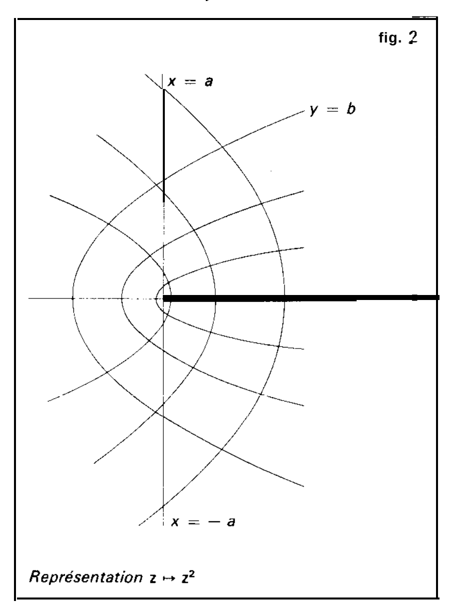
on voit que l'image de la demi-droite $x = a$ ($a \neq 0$), $y > 0$ est la demi-parabole :

$$(1) \qquad X = a^2 - \frac{Y^2}{4a^2}, \quad aY > 0.$$

La parabole (1) admet l'axe réel pour axe de symétrie ; son foyer est en 0 et son sommet d'ordonnée positive. L'image de la demi-droite $x = 0$, $y > 0$ est le demi-axe réel négatif. De plus, la droite? = $b (b > 0)$ est transformée en la parabole :

$$(2) \qquad X = \frac{Y^2}{4b^2} - b^2,$$

qui a pour axe l'axe réel ; son foyer est en 0 et son sommet d'ordonnée négative. Les paraboles de la famille (1) sont évidemment toutes orthogonales à celles de la famille (2) (fig. 2). Remarquons que l'application conforme considérée se prolonge continûment à la frontière $y = 0$ du domaine (en



Représentation $z \mapsto z^2$

fait $z \mapsto z^2$ est holomorphe dans le plan tout entier. mais non injective) ; l'image de cette frontière est la frontière du domaine image (le demi-axe réel positif).

Par restriction de la représentation conforme précédente, on obtient une représentation conforme du demi-plan $y > b$ sur l'extérieur d'une parabole.

On peut étudier de la même façon la fonction holomorphez $\mapsto z^n$ (n entier > 0), qui est injective dans le secteur angulaire

$$0 < \arg z < 2\pi/n,$$

et définit une représentation conforme de ce secteur angulaire, sur le plan fendu suivant l'axe réel positif. Ici encore, la représentation se prolonge par continuité à la frontière, mais, au point 0, la transformation cesse d'être conforme, car la dérivée s'annule ; chaque angle en 0 est multiplié par n dans cette transformation.

Les applications réciproques de ces représentations conformes fournissent de nouveaux exemples : dans le plan fendu suivant le demi-axe réel positif, on définit sans ambiguïté la fonction $z \mapsto z^{1/n}$ par la condition :

$$0 < \arg z^{1/n} < 2\pi/n,$$

et on obtient une représentation conforme du plan fendu sur un secteur angulaire d'amplitude $2\pi/n$. Pour tout exposant α réel > 0, on peut définir une représentation conforme $z \mapsto z^\alpha$ dans le secteur angulaire :

$$0 < \arg z < \inf(2\pi, 2\pi/\alpha),$$

en imposant la condition :

$$0 < \arg z^\alpha < 2\pi\alpha,$$

qui détermine une branche holomorphe de la fonction considérée, injective dans le secteur décrit par: ; l'image de cette représentation conforme est le secteur angulaire :

$$0 < \arg z < \inf(2\pi\alpha, 2\pi).$$

En composant de telles représentations conformes, on peut construire une représentation conforme d'un secteur angulaire quelconque sur un autre, par exemple sur le demi-plan supérieur. Pour préciser la représentation $z \mapsto z^{1/2}$ du plan fendu sur le demi-plan supérieur, décrivons les transformées des droites parallèles aux axes. L'image de la droite x = a (a $\neq$ 0) est la demi-hyperbole équilatère :

(3) $$X^2 - Y^2 = a, \quad Y > 0,$$

dont les asymptotes sont les bissectrices des axes de coordonnées ; si a > 0, les sommets sont sur l'axe réel et l'image est formée de la moitié supérieure de chacune des branches de l'hyperbole ; si, au contraire, a < 0, les sommets sont sur l'axe imaginaire et l'image est la branche supérieure de l'hyperbole. La droite $x = 0$ a pour image la réunion des deux demi-bissectrices des axes qui sont dans le demi-plan supérieur. Enfin, la droite $y = b$ ($b \neq 0$) est transformée en la branche supérieure de I'hyperbole équilatère :

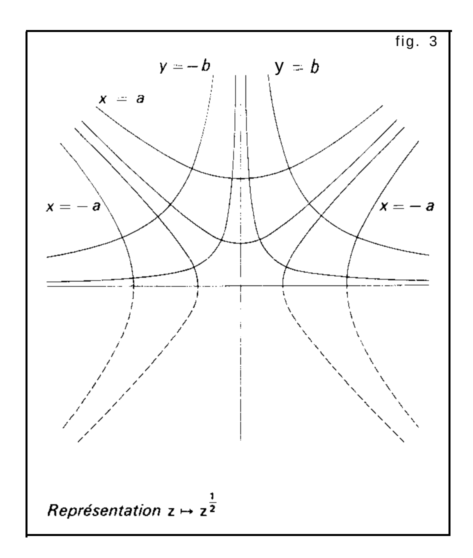
(4) $$2XY = b,$$

dont les asymptotes sont les axes de coordonnées et la demi-droite $y = 0$, $x < 0$ est transformée en la demi-droite X = 0, Y > 0. Chaque hyperbole de la famille (4) est orthogonale à toutes les hyperboles de la famille (3) (fig. 3). Le demi-plan $y > b$ est représenté conformément sur l'intérieur d'une branche d'hyperbole.

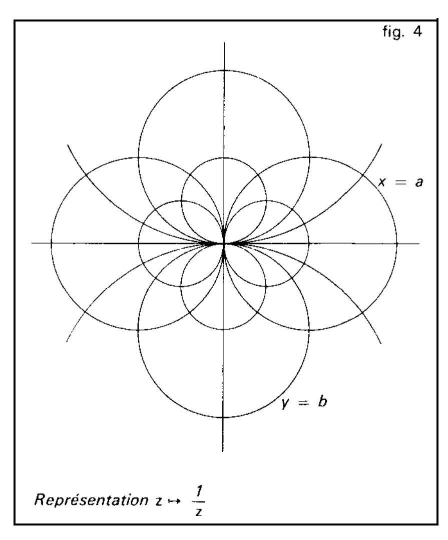
Étudions maintenant la représentation conforme définie par la fonction $z \mapsto 1/z$, qui est holomorphe et injective dans C $-$ {0} et admet pour image C {0} ; cette transformation est sa propre réciproque (transformation involutive).

Ici :

$$\frac{1}{x+iy} = \frac{x-iy}{x^2+y^2}$$



Représentation $z \mapsto z^{\frac{1}{2}}$



Représentation $z \mapsto \frac{1}{z}$

et les droites $x = a$ ($a \neq 0$) et $y = b$ ($b \neq 0$) sont respectivement transformées en les cercles passant par 0 (privés de 0) :

(5) $a(X^2 + Y^2) = X$,

(6) $b(X^2 + Y^2) = -Y$;

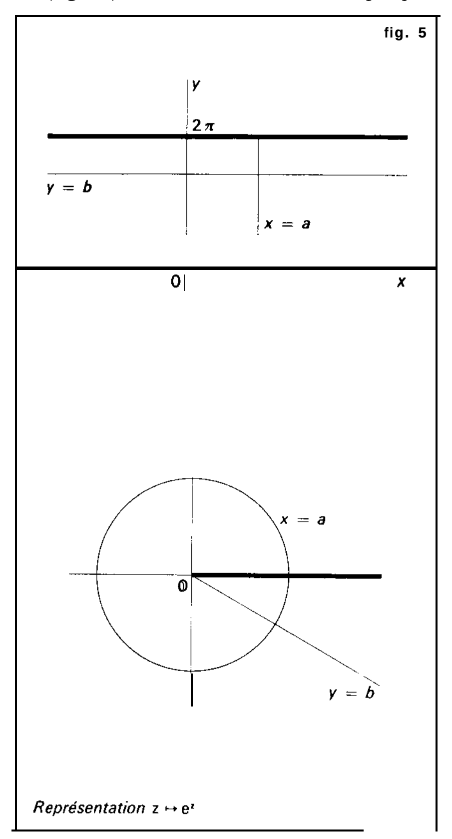
les cercles (5) sont orthogonaux aux cercles (6) ; ils sont centrés sur l'axe réel et tangents en 0 à l'axe imaginaire (fig. 4). Les droites $x = 0$ et $y = 0$ sont globalement invariantes. Plus généralement. chaque droite issue de 0 est transformée en sa symétrique par rapport à l'axe réel ; en composant avec la symétrie d'axe R, on obtient une transformation $z \mapsto 1/\bar{z}$, appelée *inversion* (de pôle 0 et de puissance 1), qui laisse globalement invariante chaque droite issue de O ; elle transforme les angles en leurs opposés. De cette transformation se déduit par restriction une représentation conforme du demi-plan $y > b$ sur le disque $b(X^2 + Y^2) + Y < 0$; en composant à droite avec la translation $z \mapsto z + ib$ et à gauche avec la similitude $z \mapsto 1 - 2biz$, on trouve une représentation conforme :

$$z \mapsto \frac{z - ib}{z + ib}$$

du demi-plan supérieur $y > 0$ sur le *disque unité* $|z| < 1$ (on peut prendre $b = 1$) ; cette transformation se prolonge à la frontière, l'image de la droite $y = 0$ étant le cercle unité privé du point -1. Il est maintenant possible de construire une représentation conforme d'un secteur angulaire quelconque sur le disque unité, puisqu'un tel secteur se représente conformément sur le demi-plan.

La fonction $z \mapsto e^z$ donne un nouvel exemple de représentation conforme (cf. EXPONENTIELLE ET LOGARITHME). Elle est holomorphe dans tout le plan et sa restriction à la *bande* $0 < \operatorname{Im} z < 2\pi$ est injective ; l'image de cette bande est le plan fendu suivant le demi-axe réel positif. Comme $e^z = e^{x+iy} = e^x e^{iy}$, c'est-à-dire $|e^z| = e^x$ et $\arg e^z = y$, nous décrirons l'image à l'aide des coordonnées polaires r et θ ($0 < \theta < 2\pi$). Les segments x = constante de la bande sont transformés

en cercles de centre 0 (privés du point réel > 0) ; les droites $y =$ constante sont transformées en demi-droites passant par 0 (fig. 5). La transformation réciproque,



Représentation $z \mapsto e^z$

notée lg, représente le plan fendu conformément sur une bande.

La fonction :

$$z \mapsto \cos z = \frac{1}{2}(e^{iz} + e^{-iz})$$

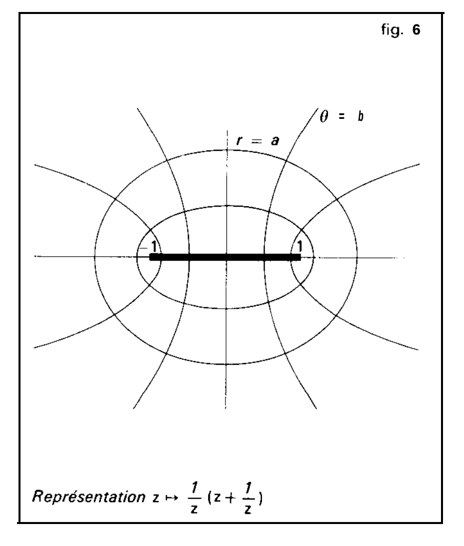
est composée des fonctions :

$$z \mapsto iz = t,$$
$$t \mapsto e^t = u,$$
$$u \mapsto \frac{1}{2}\left(u + \frac{1}{u}\right).$$

Étudions d'abord la dernière fonction : elle est holomorphe dans le plan privé de 0, et prend la même valeur aux points u et $1/u$: restreinte à l'extérieur $u > 1$ du disque unité, elle est injective et représente conformément l'extérieur du disque unité sur le plan privé de l'image du cercle unité, c'est-à-dire le plan privé du segment d'extrémité 1 et 1. Si $u = re^{i\theta}$ $(r > 1)$, les coordonnées de son image sont :

$$x = \frac{1}{2}\left(r + \frac{1}{r}\right)\cos\theta,$$
$$y = \frac{1}{2}\left(r - \frac{1}{r}\right)\sin\theta,$$

les cercles de centre 0 sont donc transformés en ellipses de foyers -1 et 1, tandis que les droites passant par 0 sont transformées en hyperboles ayant les mêmes foyers, avec dégénérescence en les axes de coordonnées (fig. 6). La même fonction

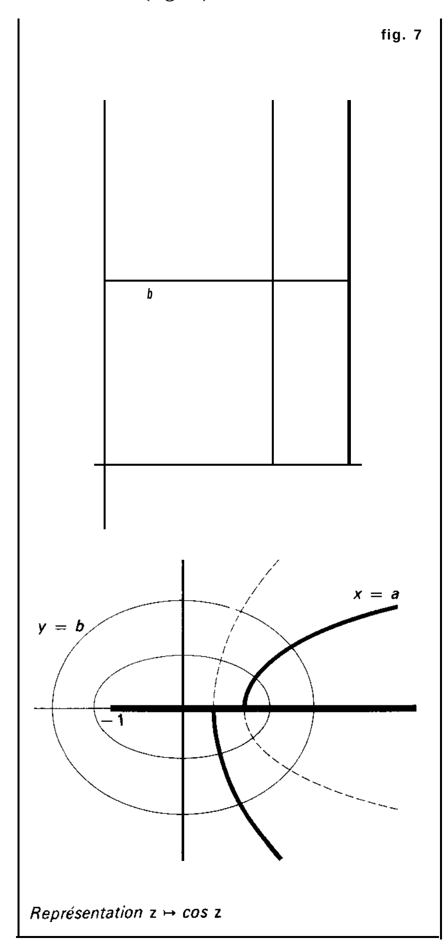


Représentation $z \mapsto \frac{1}{2}\left(z + \frac{1}{z}\right)$

donne une représentation conforme du disque unité privé de son centre sur le plan privé du segment d'extrémités 1 et 1.

Il est maintenant facile de voir que la fonction cos définit une représentation conforme de la demi-bande $0 < \operatorname{Re} z < 2\pi$, $\operatorname{Im} z > 0$ sur le plan

privé de la demi-droite $\operatorname{Im} z = 0$, $\operatorname{Re} z > -1$ (fig. 7).



Représentation $z \mapsto \cos z$

2. Le problème de la représentation conforme

Étant donné des domaines D et D' du plan, sont-ils conformément équivalents ? Dans l'affirmative, il s'agira de construire, au moins d'une manière approchée, une représentation conforme de D sur D'. Ce problème a des applications en diverses questions de physique (par exemple en hydrodynamique), car il permet de résoudre certains problèmes de Dirichlet : pour trouver une fonction harmonique u, connaissant une courbe $u = a$ (constante, qui est la frontière d'un domaine D conformément équivalent au demi-plan supérieur, on utilise une représentation conformefde D sur le demi-plan supérieur ; si f se prolonge par continuité à la frontière de D et transforme cette frontière en celle du demi-plan, soit la droite $\operatorname{Im} z = 0$, la solution est $u = \operatorname{Im}(f + a)$ (Cf. POTENTIEL ET FONCTIONS HARMONIQUES).

Si les domaines D et D' sont conformément équivalents, ils sont *homéomorphes*, c'est-à-dire qu'il existe une bijection continue de D sur D' dont la réciproque est aussi continue. Ainsi est réalisée une condition nécessaire d'isomorphisme ; mais cette condition n'est pas suffisante, car le plan C et le disque unité sont homéomorphes (l'application $z \mapsto z/(1 + |z|)$ est un homéomorphisme du premier sur le second), mais certainement pas isomorphes, puisque la fonction $z \mapsto z$ est holomorphe et bornée dans le disque unité, alors que toute fonction holomorphe et bornée dans C est constante d'après le théorème de Liouville (cf. la partie A ci-dessus - Fonctions analytiques d'une variable complexe).

D'ailleurs, comme on l'a vu ci-dessus, une grande variété de domaines sont conformément équivalents au disque unité : le demi-plan, un secteur angulaire, une bande ou une demi-bande, l'extérieur d'une parabole. En fait, Riemann a obtenu (par une démonstration un peu incomplète) le remarquable résultat suivant : Tout domaine D différent du plan C et *simplement connexe* (c'est-à-dire que tout lacet de D peut se déformer continûment dans D en un point) est conformément équivaient au disque unité.

Ce théorème a été complètement démontré par W. F. Osgood. puis par

P. Koebe, qui l'a généralisé en donnant aussi des modèles pour les domaines non simplement connexes à l'aide du disque unité privé d'un certain nombre d'arcs de cercles de centre 0. Voici les étapes de la démonstration :

a) On commence par se ramener au cas où D est borné en construisant une fonction holomorphe bornée et injective dans D (c'est assez facile). Il est alors possible de trouver des représentations conformes de D sur des domaines contenus dans le disque unité (à l'aide de similitudes, par exemple).

b) En choisissant un point a de D et en considérant l'ensemble F des représentations conformes de D sur des parties du disque unité qui transforment a en 0, on démontre que, pour un élément f de F, les propriétés suivantes sont équivalentes : (I) L'image de D par f est le disque unité. (II) $|f'(a)|$ est maximale parmi les valeurs que ce nombre peut prendre lorsque f parcourt F.

c) Il reste à démontrer l'existence d'un élément f de F qui réalise le maximum de $|f'(a)|$. Cela résulte du fait que F est un ensemble *compact* pour la « topologie de la convergence compacte » dans D et que $f \mapsto |f'(a)|$ est une fonction numérique continue dans F.

En général, on ne peut pas déterminer explicitement une représentation conforme de D sur le disque unité, mais seulement chercher à construire des approximations d'une telle représentation ; c'est un problème d'analyse numérique qui peut être difficile. La méthode de H. A. Schwarz donne explicitement une représentation conforme du demi-plan supérieur sur un polygone convexe arbitraire par une formule du type :

$$z \mapsto \int_0^z \frac{dt}{(t-a_1)^{\alpha_1} \dots (t-a_n)^{\alpha_n}}$$

où les nombres a_i sont réels et les exposants α_i compris entre 0 et 1 et de somme 2 ; pour $n = 4$ et $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/2$, l'intégrale considérée est une intégrale elliptique et donne une représentation conforme du demi-plan sur un rectangle (cf. la partie B ci-dessus - Fonctions elliptiques et modulaire).

Il est possible de déterminer toutes les représentations conformes du plan sur lui-même ou du disque unité sur lui-même. Dans le cas du plan, une représentation conforme $f : \mathbf{C} \to \mathbf{C}$ est une fonction entière qui est injective ; cela entraîne d'abord que f est un polynôme, sinon la fonction $u \mapsto f(1/u)$ aurait une singularité essentielle à l'origine et transformerait le disque unité (privé de 0) en un ensemble partout dense dans C d'après un théorème de Weierstrass ; ainsi, l'image par f de l'extérieur du disque unité serait partout dense et, par suite, rencontrerait l'image du disque unité, qui est un ouvert : c'est impossible si f est injective. De plus, f doit être de degré 1, car un polynôme de degré n a n racines ; donc f est de la forme $f(z) = az + b$ $(a \neq 0)$ et la représentation conforme est une similitude. Il est remarquable que les transformations du plan en lui-même qui conservent les angles conservent la distance euclidienne à un facteur près ; ces transformations forment un groupe à quatre paramètres, opérant transitivement.

Passons au cas du disque unité, en étudiant d'abord les automorphismes laissant fixe le point 0. Un tel automorphisme est en particulier une fonction holomorphe f telle que $f(0) = 0$ et $|f(z)| < 1$ pour tout point z du disque unité ; le lemme de Schwarz lui est applicable : $|f(z)| \leqslant |z|$, avec égalité seulement si $f(z)$ est proportionnel à z (cf. la partie A ci-dessus - Fonctions analytiques d'une variable complexe) ; le même lemme

appliqué à f^{-1} donne $|z| \leqslant |f(z)|$, donc l'égalité a lieu et $f(z) = az$ avec une constante a de module 1. Si f est un automorphisme quelconque du disque unité, on pose $b = f^{-1}(0)$. L'application :

$$g : z \mapsto \frac{z + b}{1 + \bar{b}z}$$

est une représentation conforme du disque sur lui-même qui transforme 0 en b. Donc $f \circ g$, qui est conforme et laisse 0 fixe, est une rotation $z \mapsto az$ ($|a| = 1$), et :

$$f(z) = a\,\frac{z + b}{1 + \bar{b}z}$$

Les automorphismes du disque unité sont ainsi les transformations homographiques qui le laissent invariant ; ils forment un groupe à trois paramètres transitif dans le disque unité. On peut montrer qu'il existe une métrique riemannienne dans le disque unité qui est invariante par ce groupe ; sa courbure est constante et négative de sorte que la géométrie correspondante est celle de N. I. Lobatchevsky (c'est le fameux modèle de Poincaré pour la géométrie non euclidienne).

Les résultats obtenus pour le disque unité se transposent au demi-plan. Comme le passage de l'un à l'autre s'opère au moyen d'une transformation homographique, les automorphismes du demi-plan supérieur sont les transformations homographiques qui le laissent invariant :

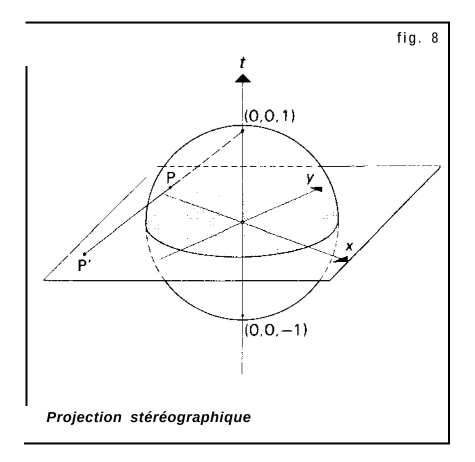
$$z \mapsto \frac{az + b}{cz + d},$$

avec a, b, c, d réels et $ad - bc > 0$.

3. Surfaces de Riemann

Projection stéréographique et sphère de Riemann

Considérons la sphère $\mathbf{S}_2$ de centre 0 et de rayon 1 dans l'espace $\mathbf{R}^3$ (où les coordonnées sont notées x, y, t). La projection stéréographique de pôle (0, 0, 1) sur le plan $t = 0$ est l'application qui, à chaque point (x, y, t) de la sphère distinct de (0, 0, 1), associe le point où la droite joignant (0, 0, 1) à (x, y, t) rencontre le plan $t = 0$ (fig. 8).



Projection stéréographique

Ainsi, l'image de (x, y, t) est le point :

$$\left(\frac{x}{1-t}, \frac{y}{1-t}\right),$$

soit :

$$\frac{x + iy}{1 - t}$$

avec la notation complexe. Il est facile de montrer que cette application conserve les angles (c'est-à-dire que l'application linéaire tangente possède cette propriété). C'est une représentation conforme de la sphère privée du pôle (0.0, 1) sur le plan C.

On peut aussi considérer la projection stéréographique de pôle (0, 0, -1), qui s'écrit :

$$(x, y, t) \mapsto \frac{x + iy}{1 + t},$$

et représente conformément la sphère privée de (0, 0, -1) sur C. Il n'a pas été

tenu compte des questions d'orientation et un même angle orienté sur la sphère est transformé en des angles opposés par les deux projections ; ce défaut se corrige en composant la deuxième projection avec la symétrie d'axe réel, ce qui donne la représentation conforme :

$$(x, y, t) \mapsto \frac{x - iy}{1 + t}$$

de la sphère privée de (0, 0, − 1) sur C. Si z et z' sont les images d'un même point (x, y, t) distinct de (0, 0, 1) et de (0, 0, − 1) par nos deux représentations, alors :

$$zz' = \frac{(x + iy)(x - iy)}{(1 - t)(1 + t)} \frac{x^2 + y^2}{1 - t^2} 1 :$$

ainsi on passe de l'une à l'autre par la transformation $z \mapsto 1/z$.

Si D est une partie ouverte de la sphère, la première projection stéréographique identifie D privé éventuellement du pôle (0, 0, 1) à un ouvert D' de C, tandis que la seconde projection identifie D privé éventuellement de (0. 0, 1) à un autre ouvert D" du plan. On dira qu'une fonction numérique complexe f définie dans D est holomorphe si les fonctions correspondantes dans D' et D" sont holomorphes ; cette définition est cohérente parce que la transformation $z \mapsto 1/z$ est un isomorphisme de D' privé éventuellement de l'origine sur D" privé éventuellement de l'origine. Avec cette notion de fonction holomorphe, la sphère S_2 s'appelle *sphère de Riemann.* Le plan s'identifie par la projection stéréographique de pôle (0, 0, 1) au complémentaire de (0, 0, 1) dans la sphère de Riemann ; comme ce point a pour image 0 par l'autre projection, il s'appellera *point à l'infini* noté ∞ (prolongeant ainsi $z \mapsto 1/z$ en posant $1/0 =$ CO). La sphère de Riemann, ainsi considérée comme C complété par un point à l'infini, peut aussi s'identifier à la droite projective complexe P,(C).

Pour déterminer les automorphismes de la sphère de Riemann, remarquons d'abord que ceux qui laissent fixe le point à l'infini donnent par restriction des automorphismes de C, c'est-à-dire des similitudes :

$$z \mapsto az + b.$$

Si f est un automorphisme quelconque, posons $c = f^{-1}(\infty)$; l'application :

$$g : z \mapsto c + 1/z$$

transforme ∞ en c et c'est un automorphisme de la sphère de Riemann, donc $f \circ g$ laisse ∞ fixe et c'est encore un automorphisme ; il en résulte que $f \circ g$ est une similitude $z \mapsto az + b$ et que :

$$f(z) = \frac{a}{z - c} + b.$$

Les représentations conformes de la sphère de Riemann sur elle-même sont donc les transformations homographiques :

$$z \mapsto \frac{az + b}{cz + d}$$

(a. b, c, d complexes tels que $ad - bc \neq 0$). Elles forment un groupe à six paramètres réels et transforment les cercles en cercles (groupe circulaire).

Courbes analytiques et surfaces de Riemann

La structure qui a été définie précédemment sur la sphère s'exprime bien dans le langage des variétés,

D'une manière générale. on appelle *variété analytique complexe* de dimension 1, ou courbe analytique complexe (régulière). ou encore, par abus de langage, surface de

Riemann, un espace topologique séparé X muni d'un atlas analytique complexe maximal à valeurs dans des ouverts de C. Cette définition repose sur la notion de carte de X à valeur dans un ouvert de C : il s'agit d'un homéomorphisme d'un ouvert de X sur un ouvert de C ; deux cartes dont les ouverts de définition se rencontrent définissent un homéomorphisme appelé changement de carte entre les deux images de l'intersection des ouverts de définition. Un atlas analytique complexe est une famille de cartes dont les ouverts de définition recouvrent X et tel que tous les changements de cartes (entre des cartes de la famille) soient holomorphes ; par exemple, la projection stéréographique de pôle (0, 0, 1) et celle de pôle (0, 0, − 1) composée avec la symétrie d'axe R sont des cartes de $\mathbf{S}_2$ et le changement de cartes est $z \mapsto 1/z$; ces cartes forment un atlas analytique complexe de $\mathbf{S}_2$ qui définit une structure de courbe analytique complexe sur la sphère (droite projective complexe). Voici encore un exemple important dans la théorie des fonctions elliptiques : étant donné deux nombres ω et ω' dont le rapport n'est pas réel, considérons le quotient X de C par le sous-groupe engendré par ω et ω' ; muni de la topologie quotient de celle de C, c'est un espace séparé (homéomorphe au tore $\mathbf{T}^2$) sur lequel on définit un atlas analytique complexe à l'aide des ouverts de C assez petits pour que l'application canonique π de C sur X y soit injective, en prenant pour cartes les réciproques des restrictions de π à de tels ouverts ; X devient ainsi une courbe analytique complexe (courbe elliptique).

Comme une fonction holomorphe est a fortiori différentiable (au sens réel), tout atlas analytique complexe est aussi un atlas différentiable et définit une structure de variété différentiable de dimension réelle 2. Une courbe analytique complexe possède donc une structure de surface différentiable ; cette surface est orientable et triangulable (la dernière propriété est due à T. Radó, 1925).

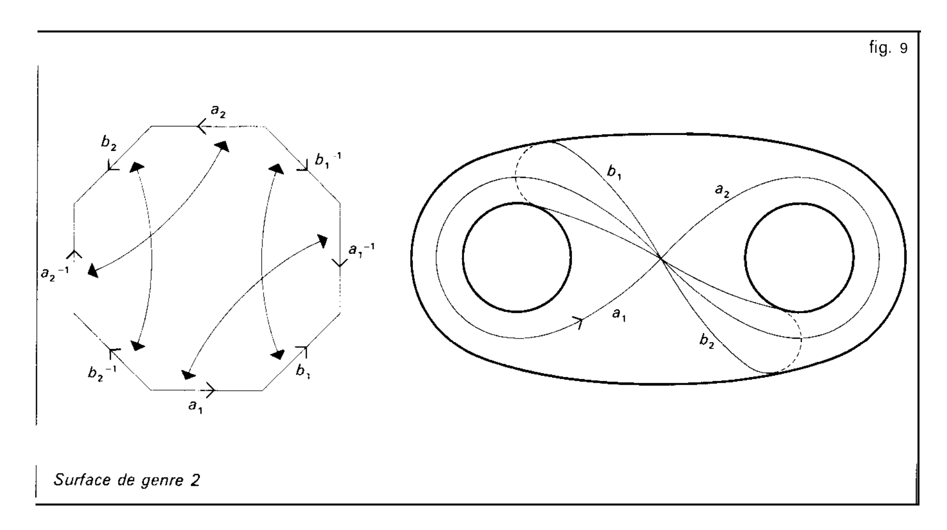
On démontre qu'une surface orientable triangulable connexe et *compacte* est homéomorphe soit à la sphère $\mathbf{S}_2$, soit à une « sphère à p anses » (ou « tore à p trous ») obtenue en identifiant deux à deux les côtés d'un polygone à 4 p côtés selon le symbole :

$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \dots a_p b_p a_p^{-1} b_p^{-1}$$

(où les $4p$ côtés sont nommés dans l'ordre où ils se présentent sur le bord orienté du polygone, a_1^{-1} devant être identifié à a_1 après avoir renversé son orientation, etc., fig. 9). Chaque côté du polygone devient un lacet sur la surface et les classes d'homotopie des 2 p lacets ainsi obtenus engendrent le groupe fondamental de cette surface ; l'entier p s'appelle le *genre* topologique de la surface. La sphère $\mathbf{S}_2$ est simplement connexe ; on lui attribue le genre 0. Le tore $\mathbf{T}^2$ est de genre 1. Ces résultats s'appliquent aux courbes analytiques complexes connexes et compactes.

Tout ouvert d'une courbe analytique complexe est muni d'une manière naturelle d'une structure induite qui en fait aussi une variété analytique complexe de dimension 1. Plus généralement, si X est une courbe analytique complexe et si f: X' $\mapsto$ X est un homéomorphisme local, l'espace X' a une structure naturelle de courbe analytique complexe provenant de celle de X. Cela s'applique au revêtement universel $\tilde{X}$ de X ; par exemple, le revêtement universel d'une courbe elliptique est C.

La notion de fonction holomorphe sur une courbe analytique complexe se définit en procédant comme ci-dessus pour la sphère de Riemann. Plus généralement, on



Surface de genre 2

peut définir la notion d'application holomorphe d'une courbe analytique complexe dans une autre.

Deux courbes analytiques sont dites isomorphes ou conformément équivalentes s'il existe un homéomorphisme holomorphe de l'une sur l'autre ; il est facile de voir que l'application réciproque d'un tel homéomorphisme est aussi holomorphe. Le théorème de représentation conforme de Riemann se généralise ainsi :

Toute courbe analytique complexe connexe et simplement connexe est isomorphe à l'une des suivantes :

a) La sphère de Riemann (droite projective complexe) ;

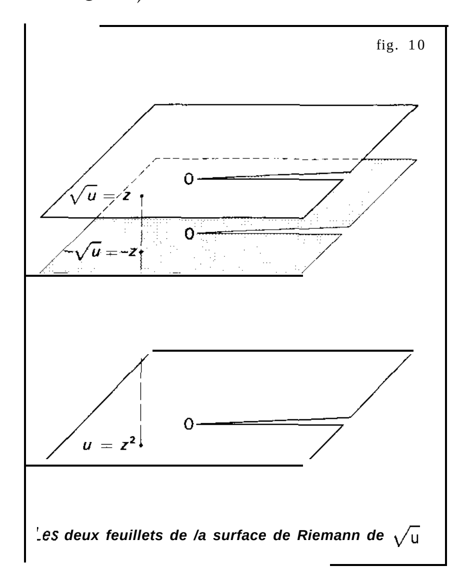
b) Le plan C (droite affine complexe) ;

c) Le disque unité (ou le demi-plan).

On distingue aisément les cas *a*, *b*, *c*, car la sphère est la seule à être compacte. et le disque unité le seul à admettre des fonctions holomorphes bornées non constantes. Le revêtement universel d'une courbe analytique complexe connexe X est isomorphe à l'une des courbes précédentes ; si c'est la sphère, la courbe X est elle-même isomorphe à la sphère ; si c'est le plan, alors X est isomorphe au plan ou au plan privé d'un point, ou bien est une courbe elliptique.

Une application holomorphe non constante $f : X \mapsto Y$ est toujours ouverte. Ses *fibres* $f^{-1}(y)$ ($y \in Y$) sont discrètes et, pour tout point $x \in X$, il existe des cartes φ et ψ de X et Y définies respectivement dans des voisinages de x et de J'(x) telles que $\varphi(x) = 0$, $\psi(f(x)) = 0$ et que $\psi \circ f \circ \varphi^{-1}$ soit de la forme $z \mapsto z^n$, où n est un entier $\geqslant 1$, que l'on appelle l'indice de ramification de f en x ; en général $n = 1$, et l'ensemble des points de X où l'indice de ramification est $\geqslant 2$ (points de ramification def') est discret. Par exemple, l'application $z \mapsto z^2$ de C dans C est ramifiée seulement en 0, avec 2 comme indice de ramification ; si on la prolonge à la sphère de Riemann en posant $\infty^2 = \infty$, on obtient une application holomorphe de la sphère de Riemann sur elle-même, ramifiée en 0 et en ∞, et dont les fibres au-dessus des points $\neq 0, \infty$ ont toutes 2 points. L'image réciproque, par cette application, du plan fendu C-R+, se compose de 2 « feuillets » isomorphes au plan fendu (ils

correspondent aux 2 déterminations de $z^{1/2}$, **fig.** 10).



Les deux feuillets de la surface de Riemann de $\sqrt{u}$

On appelle *surface de Riemann* une courbe analytique *connexe* X munie d'une application holomorphe non constante à valeurs dans la sphère de Riemann ; ainsi (S_2, $z \mapsto z^2$) est une surface de Riemann, grâce à laquelle il est possible de prolonger analytiquement une branche holomorphe de la racine carrée : une telle branche est définie initialement dans le plan fendu, mais on peut la considérer comme définie dans l'un des 2 feuillets de la surface de Riemann (c'est alors l'application identique de la sphère de Riemann), ce qui permet de définir un prolongement analytique sur la surface de Riemann. D'une manière générale, le prolongement analytique d'une fonction holomorphe définie dans un ouvert de C conduit à construire une surface de Riemann sur laquelle on peut définir ce prolongement. Ainsi la fonction ln est définie sur la surface de Riemann (C, $z \mapsto e^z$), dont l'image est C $-$ {0} ; cette surface de Riemann n'est ramifiée en aucun point et a une infinité de feuillets. La surface de Riemann de la fonction algébrique y de x définie par l'équation :

$$y^2 = 4x^3 + bx + c$$

(avec $b^3 + 27c^2 \neq 0$) est la courbe d'équation homogène :

$$y^2t = 4x^3 + bxt^2 + ct^3$$

dans le plan projectif complexe $P_2(C)$, munie de l'application dans la sphère de Riemann $P_1(C)$, qui transforme le point (x, y, t) en (x, t) pour $t \neq 0$ et le point (0, 1, 0) en (1,0) = ∞ (coordonnées homogènes) ; elle a 2 feuillets et est ramifiée au-dessus des solutions de $4x^3 + bx + c = 0$; en paramétrant la courbe à l'aide des fonctions elliptiques $x = \wp(u)$, $y = \wp'(u)$ ($u \in C$), on voit qu'elle est isomorphe au quotient de C par un sous-groupe discret de rang 2 (courbe elliptique, cf. la partie A ci-dessus Fonctions elliptiques et modulaire).

Riemann a démontré que toute surface de Riemann compacte est la surface de Riemann d'une fonction algébrique. Autrement dit, **toute** courbe analytique complexe compacte est isomorphe à une courbe algébrique d'un certain espace projectif complexe (cf. COURBES ALGÉBRIQUES). Il est remarquable que le genre topologique de la courbe analytique considérée est égal au genre algébrique de la courbe algébrique isomorphe ; dans l'isomorphisme, les fonctions ou les formes différentielles méromorphes sur la courbe analytique s'identifient aux fonctions ou aux formes différentielles rationnelles sur la courbe algébrique.

CHRISTIAN HOUZEL

Bibliographie

L. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, 1963 / H. BEHNKE & F. SOMMER, *Theorie der analytischen Funktionen einer komplexen Veränder-*

lichen, Springer, Berlin, 1955 / P. A. GRIFFITHS, *Introduction to Algebraic Curves*, American Mathematical Society. Providence (R. I.). 1989 / R. NEVANLMNA. *Uniformisierung*, Berlin, 1953 / H. POINCARÉ. *Œuvres*, vol. II. IV et IX, Gauthier-Villars, Paris, 1916-1954 / B. RIEMANN, *Œuvres mathématiques*, Gauthier-Villars, 1898. repr. en fac-similé. J. Gabay. Paris. 1990 / G. SPRINGER. *Introduction to Riemann Surfaces*, Chelsea Publ., New York, 2' éd. 198 l / S. STOILOW, *Leçons sur les principes topologiques de la théorie des fonctions analytiques*, Gauthier-Villars. 1938 / H. WEYL, *The Concept of a Riemann Surface*, Addison-Wesley, Londres, 1964.

FONCTIONS CONVEXES

→ CONVEXITÉ - Fonctions convexes

FONCTIONS DE BESSEL

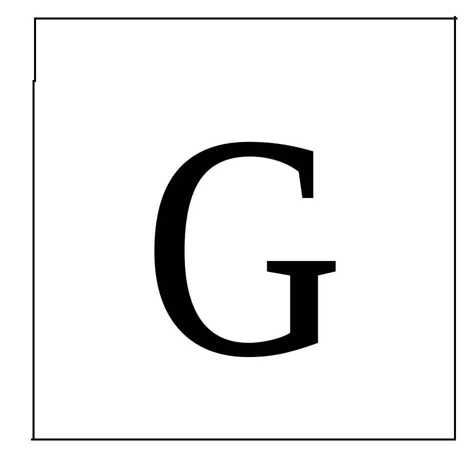
→ BESSEL FONCTIONS DE

FONCTIONS HARMONIQUES

→ POTENTIEL & FONCTIONS HARMONIQUES

FORMES QUADRATIQUES

→ QUADRATIQUES FORMES



GAMMA FONCTION

Introduites pour la première fois comme nouvelles transcendantes par L. Euler, la fonction gamma et la fonction bêta, qui s'y ramène, sont les plus importantes « fonctions spéciales » étudiées, au fur et à mesure des besoins, depuis le XVIIIe siècle. C'est ainsi que la fonction gamma intervient dans de nombreuses estimations asymptotiques des « grands nombres », en statistique notamment ; elle intervient aussi dans la théorie des séries de Dirichlet (cf. théorie des NOMBRES Théorie analytique ; fonction ZÊTA).

Nous avons choisi ici d'aborder la fonction gamma dans le domaine réel. Appliquant le principe du prolongement analytique (cf. FONCTIONS ANALYTIQUES Fonctions analytiques d'une variable complexe, chap. 1), on obtient ensuite l'extension au champ complexe de la plupart des formules.

La fonction gamma dans le domaine réel

Une intégration par parties montre facilement que, pour tout entier positif n, on a :

(1) $\int_0^\infty e^{-t}t^n dt = 1.2.3 \ldots (n-1)n = n\,!$;

mais l'intégrale (1) garde un sens pour des valeurs non nécessairement entières de n, d'où l'idée d'extrapoler ainsi la suite des factorielles. On pose traditionnellement :

(2) $$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$$

(intégrale eulérienne de seconde espèce), forme due à Euler (1781). Pour $x > 0$, cette intégrale est convergente au voisinage de 0, car $e^{-t}t^{x-1} \sim t^{x-1}$ pour x tendant vers 0, avec $x - 1 > -1$; la convergence pour l'infini résulte de la présence du terme exponentiel e^{-t}. En fait, on peut montrer que l'intégrale (2) et toutes les intégrales obtenues en dérivant un nombre quelconque de fois par rapport à x sous le signe d'intégration sont uniformément convergentes au voisinage de 0 et de $+\infty$. La fonction Γ est donc indéfiniment dérivable pour $x > 0$, de dérivées :

$$\Gamma^{(n)}(x) = \int_0^\infty (\ln\ t)^n e^{-t}t^{x-1}\,dt.$$

Remarquons qu'avec la définition (2) on a, en tenant compte de (1) :

$$\Gamma(1) = 1, \quad \Gamma(n) = (n-1)\,!,$$

ce qui suggère la convention généralement adoptée $0\,! = 1$.

Relation fonctionnelle et graphe

Remplaçant x par $x + 1$ dans (2) et intégrant par parties, on obtient :

$$\int_a^A e^{-t}t^x\,dt = \Big|-e^{-t}t^x\Big|_a^A + x\int_a^A e^{-t}t^{x-1}\,dt$$

$$= e^{-a}a^x - e^{-A}A^x + x\int_a^A e^{-t}t^{x-1}\,dt,$$

ce qui donne, en faisant tendre a vers 0 et A vers l'infini, la relation fonctionnelle :

(3) $$\Gamma(x+1) = x\Gamma(x)$$

Par récurrence, on en déduit facilement :

$$\Gamma(x) = \frac{\Gamma(x+n)}{x(x+1)\ldots(x+n)};$$

cette relation permet de définir $\Gamma(x)$ pour x réel négatif, $-n < x < -n+1$. On a ainsi défini $\Gamma(x)$ pour tout nombre réel x qui n'est pas un entier négatif ou nul.

La fonction $\ln \Gamma$ est convexe sur $]0, +\infty]$; en effet, l'inégalité de Schwarz montre que :

$$\Gamma'^2 \leqslant \Gamma\Gamma'',$$

d'où $(\ln \Gamma)'' \geqslant 0$. A fortiori, la fonction Γ est convexe. Comme $\Gamma(2) = \Gamma(1) = 1$, la fonction Γ atteint son minimum sur RT en un point compris entre 1 et 2. La figure 1 représente le graphe de cette fonction.

Formules d'Euler et de Weierstrass

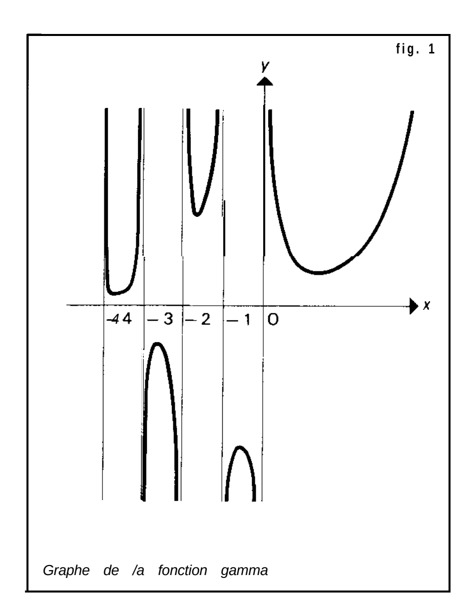
Pour n tendant vers l'infini,

$$\left(1-\frac{t}{n}\right)^n$$

tend vers e^{-t} pour tout t, et cela suggère la formule (qu'il faut, bien entendu, démontrer rigoureusement) :

$$\Gamma(x) = \lim_{n\to\infty}\int_0^n t^{x-1}\left(1-\frac{t}{n}\right)^n dt$$

$$= \lim_{n\to\infty}\left[n^x\int_0^1 u^{x-1}(1-u)^n du\right], \quad x > 0,$$



Graphe de la fonction gamma

la seconde intégrale s'obtenant en faisant le changement de variable $t = nu$ dans la première. Or, un calcul facile montre que :

$$B(x, n+1) = \int_0^1 u^{x-1}(1-u)^n du$$
$$= \frac{n!}{x(x+1)\dots(x+n)}, \quad x > 0;$$

d'où la ***formule d'Euler*** :

$$(4) \quad \Gamma(x) = \lim_{n\to\infty} \frac{n^x n!}{x(x+1)\dots(x+n)}, \quad x > 0.$$

Pour transformer cette expression, on peut écrire :

$$n^x = e^{x \ln n} = e^{x(\ln n - 1 - 1/2 - \dots 1/n)} e^{x/1} e^{x/2} \dots e^{x/n};$$

or la quantité :

$$1 + \frac{1}{2} + \dots + \frac{1}{n} - \ln n$$

tend vers une limite γ (la célèbre constante d'Euler $\gamma \sim 0{,}577\,2$) lorsque n tend vers l'infini. Divisant chacun des termes du produit $(x+1)\dots(x+n)$ par l'entier correspondant pris dans $n!$, on a donc :

$$\frac{1}{\Gamma(x)} = xe^{\gamma x} \lim_{n\to\infty} \prod_{k=1}^{n} \left(1 + \frac{x}{k}\right) e^{-x/k},$$

$$(5) \qquad = xe^{\gamma x} \prod_{k=1}^{\infty} \left(1 + \frac{x}{k}\right) e^{-x/k}, \quad x > 0,$$

puisque le produit infini est convergent (cf. SÉRIES ET PRODUITS INFINIS) ; ce développement en produit infini a été obtenu par Weierstrass.

Comportement asymptotique

Le comportement de la fonction gamma lorsque la variable x tend vers l'infini est décrit par la *formule de Stirling :*

$$(6) \quad \Gamma(x+1) \sim x^x e^{-x} \sqrt{2\pi x}, \quad x \to +\infty,$$

qui donne, en particulier, un « infiniment grand » équivalent à la factorielle :

$$n! \sim n^n e^{-n} \sqrt{2\pi n}, \quad n \to +\infty;$$

on peut d'ailleurs préciser plus étroitement le comportement asymptotique de $\Gamma(x)$ (cf. calculs ASYMPTOTIQUES).

Indiquons maintenant une formule due à Legendre pour $p = 2$ et à Gauss dans le cas général :

formule de Legendre-Gauss :

$$(7) \quad \Gamma\left(\frac{x}{p}\right)\Gamma\left(\frac{x+1}{p}\right)\dots\Gamma\left(\frac{x+p-1}{p}\right)$$
$$= (2\pi)^{(p-1)/2} p^{-x+1/2}\,\Gamma(x),$$

pour tout entier $p > 1$. Pour $p = 2$, on a donc :

$$\Gamma\left(\frac{x}{2}\right)\Gamma\left(\frac{x+1}{2}\right) = \frac{\sqrt{\pi}}{2^{x-1}}\Gamma(x).$$

Intégrales eulériennes

De nombreuses intégrales définies s'expriment au moyen de la fonction gamma. C'est ainsi que, pour les intégrales eulé-

riennes de première espèce (fonction bêta), $x > 0$ et $y > 0$:

$$(8) \qquad B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt,$$

à partir de la formule (4), Euler a établi la formule fondamentale :

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)};$$

on en déduit beaucoup d'autres résultats. Par exemple, si on effectue le changement de variable $u = \sin^2 t$, on obtient :

$$\int_0^{\pi/2} \sin^{2x-1} t \cos^{2y-1} t\,dt, = \frac{1}{2}B(x,y) = \frac{\Gamma(x)\Gamma(y)}{2\,\Gamma(x+y)}.$$

Faisant $x = y = 1/2$ dans la formule précédente, on obtient la valeur :

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},$$

qui permet, en utilisant (3), de calculer plus généralement $\Gamma(n + 1/2)$.

Extension au champ complexe

La formule de Weierstrass (5) garde un sens lorsque la variable x prend des valeurs complexes. En effet, on montre par des majorations que le produit infini de terme général $(1 + z/n)e^{-z/n} = 1 + u_n(z)$ converge normalement (cela signifie que la série de terme général $u_n(z)$ converge normalement) dans tout disque $|z| \leqslant R$. Ce produit infini définit donc une fonction de z analytique dans tout le plan complexe. Nous poserons *par définition :*

$$(9) \qquad \frac{1}{\Gamma(z)} = ze^{\gamma z}\prod_{n=1}^{\infty}\left(1+\frac{z}{n}\right)e^{-z/n};$$

cette fonction admet les points $-n$, $n \in \mathbf{N}$, pour zéros simples, et, par suite, la fonction $\Gamma(z)$ est méromorphe et ses pôles, simples, sont ces points $-n$. La formule (9) est la factorisation de Weierstrass de la fonction entière $1/\Gamma$ (cf. FONCTIONS ANALYTIQUES Fonctions analytiques d'une variable complexe, chap. 9).

Le principe du prolongement analytique permet alors de voir que de nombreuses formules établies ci-dessus pour x réel positif restent vraies pour z complexe. Par exemple la relation fonctionnelle s'écrit :

$$(10) \qquad \Gamma(z+1) = z\,\Gamma(z);$$

la formule (7) de Legendre-Gauss s'étend également. D'autre part, de (9) résulte facilement, en faisant à l'envers le calcul du chapitre 1, que l'on a pour tout z :

$$(11) \qquad \frac{1}{\Gamma(z)} = \lim_{n\to\infty}\frac{z(z+1)\dots(z+n)}{n^z n!},$$

où $n^z = e^{z\ln n}$ est fonction entière de z. Enfin, pour tout nombre complexe z de partie réelle strictement positive, on a :

$$(12) \qquad \Gamma(z) = \int_0^{\infty} e^{-t}t^{z-1}\,dt$$

l'intégrale étant uniformément convergente pour $0 < a \leqslant \operatorname{Re} z \leqslant M$.

La convergence étant normale dans (9), on obtient, en prenant la dérivée logarithmique des deux membres :

$$(13) \qquad \frac{\Gamma'(z)}{\Gamma(z)} = -\gamma - \frac{1}{z} + \sum_{n=1}^{\infty}\frac{z}{n(z+n)},$$

pour $z \neq -n$, $n \in \mathbf{N}$, la convergence étant normale sur tout compact de $\mathbf{C} - (-\mathbf{N})$.

La formule des compléments

À partir de (11) et du développement eulérien de $\sin z$ (cf. EXPONENTIELLE ET LOGARITHME, chap. 5) :

$$(14) \qquad \sin z = z\prod_{n=1}^{\infty}\left(1-\frac{z^2}{n^2\pi^2}\right),$$

on obtient l'importante « formule des compléments » due à Euler :

$$(15) \qquad \frac{1}{\Gamma(z)\Gamma(1-z)} = \frac{\sin \pi z}{\pi}, \quad z \in \mathbf{C}.$$

Appliquons, par exemple, cette formule pour $z = it$, t réel. On a alors $\Gamma(1 - it) = -it\,\Gamma(-it) = -it\,\overline{\Gamma(it)}$ d'après (10), d'où $|\Gamma(it)|^2 = \pi/t \operatorname{sh} t$.

La formule des compléments peut aussi s'obtenir directement, sans utiliser (13), à partir d'une représentation, due à Hankel, de $1/\Gamma(z)$ comme intégrale curviligne le long d'un « chemin sans fin » : cette formule (16) sert d'ailleurs dans de nombreuses questions relatives à la fonction gamma.

Désignons par U l'ensemble des nombres complexes privé des réels négatifs ou nuls. Pour $z \in \mathbf{C}$ et $u \in U$, u^z désignera la détermination principale de cette fonction dans U, soit $u^z = e^{z \ln u}$, où $\ln u$ représente la détermination principale du logarithme dans U, réelle pour u réel positif. Soit alors $L : \mathbf{R} \to U$ un chemin sans fin dans U défini par $L(t) = r(t)e^{iv(t)}$, où $r(t)$ tend vers l'infini lorsque t tend vers $\pm\infty$, on suppose qu'il existe $\varepsilon > 0$ tel que :

$$\frac{\pi}{2} + \varepsilon \leqslant v(t) < \pi \quad \text{et} \quad -\pi < v(t) \leqslant -\frac{\pi}{2} - \varepsilon$$

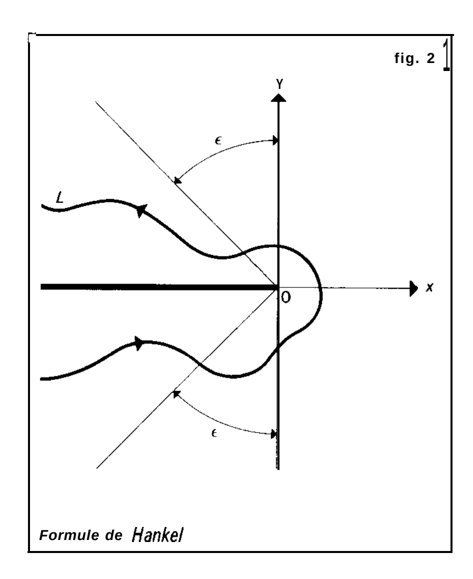
au voisinage de $+\infty$ et de $-\infty$ respectivement (fig. 2). On a alors :

$$(16) \qquad \frac{1}{\Gamma(z)} = \frac{1}{2i\pi} \int_L u^{-z} e^u \, du.$$

Interprétation par la théorie des groupes

Le corps R des nombres réels est localement compact et les caractères du groupe additif R (cf. analyse HARMONIQUE, chap. 4) sont de la forme :

$$t \mapsto e^{-ut}, \quad u \in \mathbf{C}.$$



Formule de Hankel

La composante connexe du groupe multiplicatif du corps R est le groupe $\mathbf{R}_+^*$, dont la mesure invariante est dt/t. Les caractères du groupe multiplicatif sont de la forme :

$$t \mapsto t^s, \quad s \in \mathbf{C}.$$

Si on cherche à décomposer un caractère additif selon les caractères du groupe multiplicatif, on est conduit à étudier l'intégrale sur $\mathbf{R}_+^*$ (transformée de Laplace, appelée aussi transformée de Mellin) :

$$\int_0^\infty e^{-ut} t^s \frac{dt}{t},$$

qui converge pour Re $s > 0$ et Re $u > 0$. Le changement de variable $ut = x$ joint à une intégration dans le champ complexe suivant le contour indiqué dans la fig. 2, suivi d'un passage à la limite pour $\varepsilon \to 0$ et $R \to +\infty$ (on applique le théorème de Cauchy), conduit à la relation :

$$(17) \qquad \int_0^\infty e^{-ut} t^s \frac{dt}{t} = \frac{1}{u^s} \Gamma(s),$$

où $u^s = \exp(s \ln u)$, en désignant par $\ln u$ la détermination principale du logarithme.

Le point de vue précédent montre l'analogie entre la fonction gamma et les sommes de Gauss, en arithmétique, où on considère l'anneau fini $\mathbf{Z}/n\mathbf{Z}$ des entiers modulo n et le groupe multiplicatif G,, de ses éléments inversibles. Ces deux cas relèvent de l'analyse harmonique dans les anneaux localement compacts.

La formule (17) permet en outre d'exprimer les caractères du monoïde multiplicatif N, à savoir :

$$n \mapsto \frac{1}{n^s}, \quad s \in \mathbf{C},$$

à l'aide des caractères $t \mapsto e^{-nt}$ du groupe additif R par la formule :

$$(17') \qquad \frac{1}{n^s} = \frac{1}{\Gamma(s)} \int_0^\infty e^{-nt} t^s \frac{dt}{t}$$

On en déduit qu'une série de Dirichlet peut s'écrire comme transformée de Mellin d'une série entière :

$$(18) \quad \sum_{n=1}^{\infty} \frac{a_n}{n^s} = \frac{1}{\Gamma(s)} \int_0^\infty \left(\sum_{n=1}^{\infty} a_n e^{-nt} \right) t^s \frac{dt}{t}.$$

La fonction gamma permet ainsi de ramener certains problèmes d'arithmétique multiplicative à des problèmes additifs. En particulier, la célèbre fonction zêta, intervenant dans la théorie des nombres premiers, peut s'écrire sous la forme :

$$(19) \qquad \zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{t^{s-1}}{e^t - 1} dt,$$

qui est à la base de la théorie de Riemann.

JEAN-LUC VERLEY

Bibliographie

E. ARTIN, *The Gamma Function*, Holt, New York, 1964 / N. NIELSEN, *Die Gammafunktion*, Chelsea Publ., New York, 1965 / E. WHITTAKER & G. N. WATSON, *A Course of Modern Analysis*, rééd., Cambridge Univ. Press, (Mass.), 1969.

GÉOMÉTRIE

La géométrie est communément définie comme la science des figures de l'espace. Cette définition un peu incertaine risque de conduire à inclure dans la géométrie des questions qui ne sont géométriques que dans leur langage, mais relèvent en fait d'autres domaines. Tel est le cas de l'algèbre géométrique des Grecs qui parlait du « rectangle » de deux segments pour qualifier le produit de deux nombres. Jusqu'au début des Temps modernes, presque toute la mathématique s'exprimait géométriquement : ainsi la *Géométrie* de Descartes traite non seulement de géométrie, mais aussi des équations algébriques. Et, au XIX[e] siècle, les mathématiciens étaient encore bien souvent qualifiés de géomètres, même quand ils étaient de purs analystes ou algébristes.

Plus délicat, en revanche, est le cas des domaines mixtes où des questions au départ incontestablement géométriques apparaissent très vite ne constituer qu'un chapitre de l'algèbre ou de l'analyse, et ne pouvoir être correctement traitées que par les moyens de ces disciplines. Ainsi se présentent le calcul des surfaces, le calcul des volumes, la détermination des tangentes à une courbe, et, plus généralement, l'ensemble de la géométrie infinitésimale. Historiquement, ces questions relevèrent de la géométrie pure, mais leur caractère abstrait devait bientôt se dégager et être retenu comme premier. Pourtant, ces deux modes d'approche sont trop intimement liés pour que l'on puisse songer à les séparer ; en outre, par son caractère infinitésimal, ce domaine se distingue assez nettement des autres branches de la géométrie qui sont à peu près exclusivement de caractère « fini ». Aussi n'en sera-t-il pas

question, sans que soit oubliée pour autant l'existence d'une géométrie infinitésimale « directe » où ont excellé, au XVII[e] siècle, Pierre de Fermat et Pascal, au XVIII[e] siècle, Jean-Baptiste Meusnier de La Place, au XIX[e] siècle, Charles Dupin, et qui, même au XX[e] siècle, notamment avec les travaux de Georges Bouligand, demeure un champ de recherches certes assez limité, mais digne d'attention.

Pour des raisons similaires on ne retiendra ni la géométrie algébrique ni la trigonométrie, bien que le point de vue infinitésimal n'y soit pas aussi dominant,

Ainsi délimitée, la géométrie a un objectif fondamental assez homogène qui est l'étude des figures au sens le plus large, bien qu'elle soit fort diverse dans ses méthodes et dans ses points de vue.

Depuis Descartes, la géométrie s'est développée dans deux directions nettement distinctes : la géométrie analytique et la géométrie dite « pure » ou encore « synthétique ». La conception de l'une et de l'autre, ainsi que celle de leurs rapports, a connu des vicissitudes qui constituent l'un des aspects les plus intéressants de l'histoire moderne de la géométrie. On a pu notamment assister aux efforts de la géométrie pure pour sauvegarder une autonomie que menaçait sans cesse davantage le développement de l'algèbre et de l'analyse. Un de ses derniers bastions, la chaire de géométrie pure de l'École polytechnique de Paris, a été supprimé en 1956. Aujourd'hui, bien qu'elle soit encore valable à plus d'un titre, principalement en raison du rôle qu'y jouent l'imagination et l'intuition, la géométrie pure n'occupe plus qu'une place seconde dans la mathématique.

De caractère plus intrinsèque apparaît le pluralisme de la géométrie qui s'est manifesté, surtout depuis le début du XIX[e] siècle, avec la constitution de la géométrie projective en discipline autonome et avec la naissance des géométries non euclidiennes. Diversité qui fut pleinement comprise et dominée par le mathématicien allemand Félix Klein, dans son célèbre programme d'Erlangen.

La diversification devait s'accentuer lorsque, se libérant plus nettement encore de l'espace physique euclidien où elle était demeurée enfermée depuis plus de deux millénaires, la géométrie, d'une part, s'est étendue aux espaces à plus de trois dimensions et, d'autre part, a intégré systématiquement les éléments imaginaires, Dans une étape ultérieure, la géométrie devait « éclater » pour s'insérer dans les deux grandes structures générales de la mathématique moderne, l'algèbre et la topologie.

Dès lors, on peut comprendre les vraies raisons de l'incertitude qui a pesé, tout au long de l'histoire de la géométrie, sur sa nature et sur ses rapports avec les autres domaines de la mathématique. On comprend mieux également pourquoi, très tôt, elle s'est sentie menacée dans son autonomie par le développement de l'algèbre et de l'analyse, et pourquoi, finalement, en dépit du constant effort d'unification qui marque son histoire, notamment avec Euclide, Apollonios, René Descartes, Gérard Desargues, Jean Victor Poncelet, Michel Chasles, elle était condamnée à disparaître comme discipline autonome, pour ne plus être qu'une « illustration » des structures abstraites de la mathématique moderne.

Du très grand nombre de questions et de problèmes qui se rencontrent dans la géométrie, on retiendra essentiellement les fondements, les principes, les notions majeures et les théorèmes les plus importants.

1. La géométrie classique

La synthèse euclidienne

On rencontre déjà en Égypte ancienne, à côté d'une pratique géométrique, un début de science géométrique, comprenant notamment diverses propositions sur les propriétés du triangle et du cercle. Plus tard, en Grèce, principalement avec Thalès au VIe siècle avant J.-C., Pythagore et Hippocrate de Khios au Ve siècle, Eudoxe au IVe siècle, un nombre appréciable de résultats géométriques sont obtenus : inscription de sphères dans un cône, similitude des triangles, principales propriétés du cercle, polygones et polyèdres réguliers, sections coniques. Utilisant ces données et les complétant, Euclide (fin du IVe siècle av. J.-C.) réalise avec ses *Éléments* la première synthèse de la géométrie.

En fait, les *Éléments* comportent, à côté de la géométrie proprement dite, d'importants chapitres qui n'en relèvent aucunement (nombres entiers, nombres irrationnels, proportions, équations du premier et du second degré), ou qui n'en relèvent pas au sens restreint retenu ici pour la géométrie, savoir la détermination des surfaces et des volumes.

Outre leur caractère organique. les *Éléments* ont ceci de remarquable que leur auteur a le souci de « fonder » la géométrie : ils débutent par une série d'énoncés de base, à partir desquels sont déduites toutes les autres propositions. Une telle innovation procède essentiellement des préoccupations et de l'œuvre logique d'Aristote.

Ces énoncés se répartissent en trois catégories : des définitions (point, ligne droite, surface, angles...), des vérités considérées comme évidentes, et, de ce fait, n'appelant pas de démonstration (deux grandeurs égales à une même grandeur sont égaies entre elles, le tout est plus grand que la partie...), des demandes ou postulats, vérités non évidentes par elles-mêmes, que l'on ne sait pas démontrer, mais dont on a besoin du fait que les théorèmes que l'on en déduit apparaissent vérifiés concrètement.

La plus importante et la plus célèbre de ces demandes est le postulat des parallèles qui affirme - à la formulation près - que, par un point situé hors d'une droite, on peut mener une droite et une seule qui ne la rencontre pas, cette droite étant dite parallèle.

Une autre demande de grande portée, que l'on trouve seulement dans le corps de l'ouvrage, est le postulat dit d'Archimède, qualifié aujourd'hui de postulat de continuité : Deux points A et B étant donnés sur une droite, si, à partir de A, l'on met à la suite des segments de même longueur, on dépassera le point B après une série finie de telles opérations, si petite que soit cette longueur.

Il existe en réalité une certaine hésitation chez Euclide et ses successeurs quant à la nature exacte de ces « demandes ». Les uns estiment qu'elles sont suffisamment évidentes pour n'avoir pas à être démontrées. D'autres, au contraire, pensent que l'on doit pouvoir les démontrer et que la géométrie ne sera vraiment satisfaisante que lorsque l'on y sera parvenu. C'est cette dernière opinion qui l'emporta, donnant lieu aux tentatives infructueuses de démonstration qui occupèrent tant de géomètres jusqu'à la naissance des géométries non euclidiennes.

En dehors du postulat des parallèles, la géométrie euclidienne a été longtemps considérée comme le modèle même d'une

connaissance vraie et rigoureuse. Aujourd'hui, ses fondements se révèlent, à bien des égards, très peu assurés. Les définitions d'Euclide ne sont pas de vraies définitions, mais plutôt des descriptions d'intuitions ; elles utilisent des concepts qui sont considérés comme premiers, alors qu'ils demandent eux-mêmes à être définis. Tel est le cas des définitions du point comme ce qui n'a pas de partie, de la ligne comme longueur sans largeur, de la ligne droite comme celle qui est située semblablement par rapport à tous ses points, de l'angle plan comme l'inclinaison mutuelle de deux lignes qui se rencontrent dans un plan et qui ont des directions différentes.

C'est seulement à la fin du XIXe siècle, surtout avec David Hilbert (1862-1943), dans *Principes fondamentaux de lu géométrie (Grundlagen der Geometrie,* 1899) que la géométrie euclidienne est fondée de façon satisfaisante par une démarche dégagée de toute intuition sensible, et Hilbert démontre préoccupation étrangère à Euclide - que les axiomes retenus sont indépendants et qu'ils ne conduisent pas à des contradictions. Ensuite, des mathématiciens, tel Gustave Choquet dans *L'Enseignement de lu géométrie* (1964), ont exposé la géométrie euclidienne sous une forme élémentaire mais rigoureuse.

La présentation moderne des axiomes de la géométrie euclidienne offre, en plus de sa rigueur logique, la supériorité sur celle d'Euclide de faire apparaître ce que Georges Bouligand a appelé la « causalité » des propositions : on est en mesure de désigner les axiomes qu'utilise une proposition donnée, certaines propositions ne faisant pas appel à l'ensemble des axiomes. Ainsi, en abandonnant le postulat des parallèles on obtient des propositions qui sont valables aussi dans les géométries non euclidiennes, et l'on peut construire des géométries « non archimédiennes » où le postulat d'Archimède n'est pas vérifié. C'est dans cette perspective que se situe la conception des géométries subordonnées de Klein dont il est question plus loin.

La sphère

La géométrie de la sphère, principalement pour les besoins de l'astronomie, devait être particulièrement développée dès l'Antiquité grecque. Elle constitue, jusqu'au début des Temps modernes, un savoir assez autonome par rapport aux autres aspects de la géométrie. Très tôt, même avant les Grecs, furent étudiés dans le cercle et la sphère les rapports entre les cordes et les angles. Dans les *Sphériques*, Menelaos d'Alexandrie (Ier siècle apr. J.-C.) démontre un important théorème, valable non seulement pour les triangles sphériques, mais aussi pour les triangles plans : Si un triangle sphérique (resp. un plan) est coupé par un grand cercle (resp. une droite), les trois points d'intersection L, M, N, sont tels que les produits des sinus des arcs (resp. les segments) sans points communs sont égaux aux produits des trois autres. *À* partir de ces résultats prendra naissance la trigonométrie plane et sphérique.

Les « Coniques » d'Apollonios

Avec les *Éléments* d'Euclide et les écrits d'Archimède, les *Coniques* d'Apollonios de Perge (fin du IIIe siècle av. J.-C.) constituent l'un des ouvrages les plus complets et les plus remarquables qu'a légués la mathématique grecque. Apollonios y présente en un ensemble organique des notions et des résultats pour une part notable antérieurs à lui, mais auxquels il a joint d'importantes contributions. Apollonios unifie la définition des coniques : au lieu d'utiliser pour chaque catégorie de

coniques un cône à base circulaire différent (obtusangle pour l'hyperbole, droit pour la parabole, acutangle pour l'ellipse), il ne fait appel qu'à un seul cône à base circulaire. On lui doit aussi la dénomination des trois types de coniques. Alors qu'avant lui on ne prenait en considération qu'une seule branche de I'hyperbole, il définit l'hyperbole comme constituée de deux branches. Retenons aussi parmi les apports originaux d'Apollonios son étude des conditions d'égalité et de similitude des coniques et de la disposition de ces courbes sur un cône donné. En dehors des éléments nouveaux sur les coniques qu'apportera la géométrie projective, l'ouvrage d'Apollonios diffère peu mis à part la prolixité des explications, due surtout à l'absence de notations de l'enseignement sur les coniques qu'on dispensait en France, dans les années soixante, en classe de mathématiques élémentaires.

2. La géométrie analytique

Origines

Il est assez habituel de considérer que la géométrie analytique a été créée par Descartes. En réalité, cette vue est trop simple. Si la « géométrie analytique » est prise au sens moderne de l'expression, celle de Descartes en était encore assez éloignée. D'autre part, plusieurs éléments caractéristiques de la géométrie analytique avaient été formulés avant Descartes.

La géométrie analytique paraît consister dans l'association de trois facteurs : l'expression d'une réalité géométrique par une relation entre des quantités variables, l'usage des coordonnées, le principe de la représentation graphique. Or, si chacun de ces trois facteurs se rencontre assez tôt dans le développement de la géométrie avant Descartes, ils n'ont cependant pas été rapprochés.

Dès la plus haute antiquité, l'observation astronomique avait conduit à repérer les directions dans l'espace par deux coordonnées angulaires : hauteur au-dessus de l'horizon, écart par rapport au méridien. Et, très tôt, furent mises en évidence des relations entre ces coordonnées. Mais il s'agissait là de pratiques qui étaient à peu près sans rapport avec la science géométrique. Au contraire, c'est au cœur même de la géométrie que l'on voit intervenir chez les Grecs un calcul portant sur deux variables en vue de caractériser des réalités géométriques et d'en établir les propriétés. Chez Archimède et surtout chez Apollonios, un tel calcul est développé systématiquement pour l'étude des coniques. Apollonios écrit explicitement les équations des coniques en coordonnées obliques ayant pour origine un point de la conique et pour directions le diamètre correspondant à ce point et son diamètre conjugué :

$$y^2 = 2px + \frac{p}{a}x^2,$$

$$y^2 = 2px - \frac{p}{a}x^2,$$

$$y^2 = 2px,$$

pour l'hyperbole, l'ellipse et la parabole respectivement.

Dans une perspective tout à fait différente. Nicolas Oresme, au XIVe siècle, imagine une représentation graphique de certains phénomènes. Il distingue une *latitudo* et une *longitudo* qui correspondent à l'abscisse et à l'ordonnée d'une représentation en coordonnées rectangulaires. Cette façon de faire est inverse de celle des Grecs, puisque Oresme ne part pas d'une réalité géométrique mais exprime sous forme géométrique une relation entre des grandeurs. La conception même d'une

telle correspondance doit être considérée comme s'inscrivant dans le cadre des idées qui sont à la base de la géométrie analytique. Toutefois, les vues d'Oresme, en dépit de la grande faveur qu'elles connurent, ne furent aucunement rapprochées des pratiques « analytiques » des Grecs dont l'Occident prit connaissance vers la fin du XVIe siècle avec la publication en latin des œuvres d'Archimède et d'Apollonios.

Descartes et Fermat

Le calcul géométrique exposé par Descartes (1596-1650) dans sa *Géométrie* (1637) ne diffère guère en son principe du calcul d'Apollonios. Il porte sur deux variables que l'on peut sans doute considérer comme constituant des coordonnées ; mais on n'y trouve pas explicités des axes de coordonnées, c'est-à-dire deux droites orientées, distinctes des lignes de la figure. Toutefois, dans quelques passages de son ouvrage, Descartes précise qu'il choisit sur une droite, distincte de la figure, un point origine ; mais il ne fait pas intervenir un autre axe de coordonnées, se contentant de choisir une direction selon laquelle est mesurée la seconde variable.

Le vrai progrès réalisé par Descartes réside en ce que, au lieu de limiter un tel calcul à l'étude d'une figure donnée, comme le faisaient les Grecs, il le pose en procédé général susceptible de permettre la création d'une infinité de courbes nouvelles Malheureusement, il limite singulièrement le champ de sa géométrie en refusant d'y recevoir les « courbes décrites par deux mouvements qui n'ont entre eux aucun rapport qu'on puisse mesurer exactement ». La formule signifie que Descartes ne reconnaît que les courbes algébriques, excluant les courbes « transcendantes », dont l'étude commençait alors à se développer (logarithme, sinus et cosinus...).

Il faut, d'autre part, noter qu'à la même époque, et même un peu avant lui, Pierre de Fermat (1601-1665) avait abouti à des conceptions fort voisines. Mais, alors que Descartes adopte des notations symboliques qui représentent les constantes et les variables par des lettres, et les puissances par des exposants, Fermat demeure attaché au langage beaucoup plus lourd de l'algèbre géométrique des Grecs.

Descartes applique avec succès sa méthode à la résolution du problème dit de Pappus : Déterminer le lieu des points tels que, étant donné quatre droites et étant considéré les distances d'un point à chaque droite sous des angles déterminés, le produit de deux distances est égal au produit des deux autres. Descartes montre aisément par le calcul que ce lieu est une conique.

La nouvelle méthode suscita dans la seconde moitié du XVIIe siècle un grand nombre de travaux, concernant surtout les courbes planes algébriques (tangente, normale, centre de courbure, point d'inflexion...).

La géométrie analytique moderne

La géométrie analytique n'acquiert pleinement les traits qui la caractérisent aujourd'hui qu'au XVIIIe siècle. Tout d'abord, alors qu'elle était demeurée limitée jusque-là au plan, la géométrie analytique est étendue à l'espace. En 1700, est écrite l'équation de la sphère ; en 173 1, Alexis Clairaut (17 13-l 765) publie une étude remarquable sur les courbes à double courbure. L'apport de Leonhard Euler (1707-I 783) est particulièrement notable : dans *Introductio in analysis infinitorum* (1748), pour la première fois, il énonce le principe de l'équivalence des deux axes,

alors que jusque-là l'axe des abscisses avait conservé, par une anomalie qui nous étonne, un rôle privilégié, et il donne une formule vraiment claire du changement de coordonnées, utilisée cependant par Van Schooten dès 1649. De plus, Euler détermine l'équation des surfaces du second degré.

La géométrie analytique ne prend cependant son essor que dans la seconde moitié du XVIIIe siècle. Dans l'esprit de ses travaux sur la mécanique analytique, Louis de Lagrange (1736-1813) souligne « avec combien de facilité et de succès la méthode algébrique peut être employée pour les questions qui paraissaient être le plus du ressort de la géométrie proprement dite et les moins propres à être traitées par le calcul ». Rompant avec la méthode cartésienne qui mêlait les procédés analytiques et géométriques, les éléments du premier ordre (droite et plan) demeurant toujours envisagés de manière géométrique, Lagrange établit autour des années 1770 les équations de la droite et du plan et inaugure l'utilisation systématique de trois axes de coordonnées.

C'est dans cet esprit que Gaspard Monge (1746-1818), à partir de 1771 et, plus systématiquement, en 1795 dans ses *Feuilles d'analyse appliquée à la géométrie*, donne à la géométrie analytique son ampleur, établissant les équations des divers types de surfaces algébriques (surfaces réglées, développables, de révolution...) et résolvant analytiquement de nombreux problèmes. On peut alors dire que la géométrie moderne est née. En 1797, Sylvestre François Lacroix (1765-1843) en rédige le premier traité, mais sans encore user du terme même de géométrie analytique, intitulant son ouvrage : *Traité de calcul différentiel et intégral*.

Le XIXe siècle apporte peu de compléments notables à la géométrie analytique proprement dite. Mais le caractère arbitraire du choix des axes de coordonnées devait conduire à l'étude des invariants dans les changements de coordonnées qui. seuls, peuvent exprimer les propriétés géométriques intrinsèques des figures. À côté des travaux d'ordre algébrique qu'elle contribua à susciter, cette étude fut un des facteurs principaux du développement, au cours du XIXe siècle, des notions de vecteur et de tenseur, dont l'utilisation allait être si féconde, non seulement en mathématique pure mais aussi dans de nombreuses applications.

3. La géométrie projective

Au sens moderne du terme, on entend par géométrie projective l'étude des propriétés des figures qui se conservent par transformation homographique. Ce point de vue général ne s'est dégagé que lentement, par élargissement de conceptions plus particulières et par une clarification qui a eu notamment à distinguer les propriétés projectives des figures de leurs propriétés métriques. La géométrie projective a joué un rôle majeur dans l'évolution de la conception de la géométrie. Elle fut le principal facteur du mouvement d'idées qui, au cours du XIX' siècle, a progressivement rapproché les diverses géométries et a donné à la notion de transformation une place centrale dans la géométrie.

Le rapport anharmonique chez les Grecs

Dans les mathématiques grecques, on ne rencontre pas à proprement parler de géométrie projective, essentiellement parce que la notion de transformation des

figures n'y apparaît pas, même pas la projection centrale que semble suggérer pourtant très naturellement la considération des cônes et de leurs intersections par des plans qui engendrent les coniques. En revanche, on trouve des notions et des théorèmes qu'on rattache maintenant à la géométrie projective. Principalement, la définition du rapport anharmonique, dit aujourd'hui birapport, de quatre points alignés A, B, C, D, soit :

$$(A, B, C, D) = \frac{\overline{CA}}{\overline{CB}} : \frac{\overline{DA}}{\overline{DB}},$$

et la démonstration de la conservation de ce rapport pour les points d'intersection de toute transversale coupant quatre droites concourantes. Les Grecs se sont plus spécialement intéressés au rapport harmonique, cas particulier où le rapport anharmonique a pour valeur 1. Ils le rencontraient notamment dans l'étude des coniques, puisque, sur une droite quelconque passant par un point donné, ce point est conjugué harmonique de l'intersection de la droite joignant les points de contact des deux tangentes menées du point, par rapport aux deux points d'intersection de la droite avec la conique. Ces considérations se trouvent dans les *Coniques* d'Apollonios et dans la *Collection mathématique* de Pappus d'Alexandrie (IIe siècle apr. J.-C.). Il faut aussi mentionner, comme se rattachant à la géométrie projective, un certain nombre de théorèmes sur les segments déterminés par des transversales à des triangles ou à des quadrilatères, principalement le théorème de Menelaos déjà énoncé.

La perspective à la Renaissance

En Europe, dès le xv' siècle, les artistes, peintres et graveurs, s'intéressent surtout à la représentation sur un plan des figures de l'espace à partir du point de vue constitué par l'oeil. Ils sont ainsi amenés à l'étude de la projection centrale, et, en particulier, à la considération du point de fuite qui représente, sur le plan des projections, le point à l'infini de droites parallèles perpendiculaires à ce plan. Il faut signaler les traités de perspective de Jean Pélerin (1505) et d'Albert Dürer (1525). Dès le XVe siècle, les architectes, tels que Filippo Brunelleschi et Leon Battista Alberti, avaient contribué au développement de la perspective pratique. Des considérations de perspective se rencontrent également dans la gnomonique (art des cadrans solaires) et dans la stéréotomie ou taille des pierres. Plus tard, la perspective est développée pour les besoins des fortifications et de la « scénographie ». Mais elle demeurait encore au début du XVIIe siècle une discipline sans rapport avec la science géométrique. C'est Desargues qui, le premier, rapprochera ces deux ordres de recherches.

Desargues et Pascal

Le Français Girard Desargues (1593-1662), ingénieur et architecte, appartient au milieu des praticiens. Il a été en rapport avec les milieux savants de l'époque. Son souci d'une rationalisation et d'une simplification de la perspective par la mise en lumière de nouvelles méthodes géométriques l'amène, en 1639, deux ans après la *Géométrie* de Descartes, à publier *Brouillon projet d'une atteinte des événements des rencontres du cône avec un plan*, petit ouvrage de quarante pages, tiré seulement à cinquante exemplaires. Rédigé de façon assez obscure, utilisant des termes nouveaux, qui, pour la plupart. ne seront pas retenus, cet opuscule est accueilli avec estime par Descartes et par Fermat ; pourtant ceux-ci n'en savent pas

reconnaître l'originalité et la portée ; il se heurte à de violentes oppositions, notamment, à celles d'auteurs de traités de perspective pratique. Seul Pascal (1623-1662) comprend vraiment Desargues. Il voit en lui «un des grands esprits de ce temps et des plus versés aux mathématiques ». Pascal s'inspire très directement des vues de Desargues dans son *Essay pour les coniques* (1640), texte de quelques pages présentant déjà le vaste programme que développera le *Traité des coniques,* achevé entre 1654 et 1658, mais qui ne fut pas publié et dont il ne reste qu'un résumé et un commentaire dus à Leibniz. Philippe de La Hire (1640-1718), dans deux traités sur les coniques (1673, 1679), reprend et systématise les idées et les résultats de Desargues et de Pascal sans y apporter toutefois des éléments notablement nouveaux.

Après La Hire et pendant près d'un siècle, la géométrie projective tombe dans l'oubli. Cela s'explique surtout par l'intérêt porté alors à la géométrie analytique promue par Descartes et au développement du calcul infinitésimal auquel Leibniz et Newton venaient de donner son plein essor.

La nouveauté de l'œuvre géométrique de Desargues réside essentiellement dans l'introduction, en géométrie, de la projection centrale ; elle permet à ce mathématicien des démonstrations « par le relief ». Ainsi, il démontre pour le cercle des propriétés relatives aux polaires et aux tangentes ; ensuite, par une projection centrale qui transforme le cercle en une conique, il les étend aux coniques. D'autre part, se fondant sur le fait que, dans une perspective, des droites parallèles se transforment en droites concourantes, Desargues complète l'espace euclidien, en définissant les points à l'infini, qui seront considérés comme de même nature que les points à distance finie. Dès lors, Desargues pouvait substituer à l'étude séparée des trois catégories de coniques, selon la manière de faire d'Apollonios, une étude unique.

On lui doit en outre l'introduction de la notion et du terme d'involution, étroitement liée à la notion de la division harmonique. Ces questions, on l'a vu, n'étaient pas ignorées des Grecs, mais elles n'avaient pas encore été étudiées systématiquement.

Au langage près, assez complexe chez lui parce que dépourvu de notations symboliques, Desargues définit l'involution sur une droite, ainsi qu'on le fait aujourd'hui, comme une correspondance symétrique et donc réciproque de deux points sur une droite. Elle s'écrit sous forme canonique $xy = k$. Tout couple de points d'une involution est conjugué harmonique des deux points doubles $x = y = \pm\sqrt{k}$. Desargues envisage seulement le cas où ces points sont réels. Il montre aussi que deux couples de points définissent une involution et il s'intéresse spécialement à l'ensemble constitué par trois couples de points en involution. Ces trois couples interviennent notamment dans un théorème relatif aux coniques, dont l'énoncé et la démonstration représentent un apport remarquable de Desargues au progrès de la théorie des coniques. Ce théorème s'énonce ainsi : Les coniques d'un faisceau ponctuel défini par quatre points, dont trois quelconques ne sont pas alignés, déterminent sur chaque droite de leur plan des couples de points en involution ; les trois coniques dégénérées du faisceau qui sont les couples de côtés opposés du quadrilatère complet défini par ces quatre points satisfaisant à cette propriété. Desargues démontre ce théorème, d'abord dans le cas du cercle en utilisant

les relations de puissances, puis pour une conique quelconque en la considérant comme section plane d'un cône de base circulaire. Il semble qu'il faille chercher l'origine de ce théorème dans des lemmes de la *Collection mathématique* de Pappus d'Alexandrie ayant trait aux relations entre segments formés sur une transversale coupée par les trois couples de côtés opposés d'un quadrilatère complet. Mais Desargues est le premier à avoir considéré les coniques dans lesquelles est inscrit ce quadrilatère.

Un autre de ses apports majeurs est le théorème sur les triangles « perspectifs » : Si, dans un même plan ou dans l'espace, deux triangles ABC et A'B'C' sont tels que les trois droites joignant respectivement les trois couples de sommets homologues AA', BB', CC' se rencontrent en un même point S, les trois points de concours des couples de droites portant les côtés homologues des deux triangles sont alignés. Et réciproquement.

Desargues démontre ce théorème, dans l'espace, par des propriétés d'incidence, et, dans le plan, par l'emploi répété du théorème de Menelaos.

Ce théorème de Desargues occupe une place fondamentale dans la structure de la géométrie projective ; ce que montre notamment le fait que sa démonstration dans le plan nécessite les postulats projectifs de l'espace comme l'a établit Hilbert dans ses *Grundlagen der Geometrie* le théorème de Menelaos utilisé par Desargues supposant ces axiomes. Hilbert a pu alors définir une géométrie projective plane « non arguésienne » où le théorème de Desargues n'est pas vérifié.

L'importance de ce théorème se manifeste aussi dans le fait qu'il est à la base de la définition de l'homologie dans le plan. Seulement entrevue par Desargues, cette définition devait jouer un rôle capital dans les travaux de Jean Victor Poncelet (1788-1867).

Soit en effet dans un plan P un triangle T projeté en un triangle T' sur un plan P' à partir d'un point 0 n'appartenant ni à P ni à P'. Rabattons le plan P sur P' par rotation autour de la droite D, intersection des deux plans P et P'. T se rabat en un triangle T". Les couples de droites joignant les côtés homologues des triangles T' et T" du plan P' se rencontrent évidemment sur D dont les points sont demeurés inchangés aussi bien dans le rabattement que dans la projection. Ces trois points étant donc alignés, il en résulte, d'après la réciproque du théorème de Desargues, que les trois droites portant les couples de sommets homologues de T' et de T" se rencontrent en un même point S. On définit ainsi une correspondance réciproque point à point et droite à droite, la construction du transformé M" d'un point courant M' s'obtenant aisément par intersection de droites, dès lors que sont donnés le sommet de l'involution S, son axe D et deux couples de points homologues, A'A", B'B" (A'A" et B'B" étant soumis à la seule condition de se rencontrer en S).

Dans ses travaux sur les coniques, Pascal reprend les méthodes et les résultats de Desargues, mais de manière beaucoup plus synthétique et plus claire. 11 y ajoute des éléments nouveaux remarquables dont le plus important est le théorème de l'hexagone inscrit dans une conique : Les trois points de concours des côtés opposés d'un hexagone inscrit dans une conique sont en ligne droite. Pascal démontre ce théorème d'abord pour le cercle et l'étend ensuite par une projection aux coniques.

Il est possible que Pascal ait entrevu le théorème dual, démontré par Charles Julien Brianchon (1783-l 864) au début du

XIXe siècle : On peut toujours trouver trois diagonales concourantes dans un hexagone circonscrit à une conique.

Pascal devait déduire du théorème de l'hexagone de nombreuses propriétés des coniques, en particulier les propriétés des pôles et polaires, considérant à cet effet le cas où, deux points de l'hexagone étant confondus, un des côtés se trouve être la tangente à la conique en ce point.

Renouveau et essor de la géométrie projective

C'est seulement à la fin du XVIIIe siècle que renaît l'intérêt pour la géométrie projective. Elle devait alors connaître un essor remarquable. Son renouveau est jalonné d'une série d'oeuvres majeures : Gaspard Monge, *Application de l'algèbre à la géométrie* (1795), *Géométrie descriptive* (1795) ; Lazare Carnot, *Géométrie de position* (1803) ; Jean Victor Poncelet, *Traité des propriétés projectives des figures* (1822) ; Michel Chasles, *Aperçu historique sur le développement des méthodes en géométrie* (1837), *Traité de géométrie supérieure* (1852) ; C. Van Staudt, *Geometrie der Lage (1847).*

Chez Monge, la géométrie analytique et la géométrie pure demeurent intimement associées. De plus, son intérêt se porte surtout sur la géométrie descriptive qui représente les figures de l'espace par deux projections orthogonales sur deux plans perpendiculaires. Monge ne semble pas avoir saisi toute l'importance de la projection centrale. 11 n'en eut pas moins une influence décisive sur le développement ultérieur de la géométrie projective, notamment par le sens de l'espace qui imprègne tout son enseignement.

Avec Poncelet s'affirme beaucoup plus nettement le dessein de constituer la géométrie projective en discipline autonome. Poncelet veut « rendre la géométrie enfin indépendante de l'analyse algébrique », et, pour lui, les propriétés projectives des figures comptant parmi les plus générales que l'on connaisse méritent à ce titre seul l'attention des géomètres. Il entend faire cesser un état de choses où « la géométrie analytique offre des moyens généraux et uniformes » alors que «jusqu'ici l'autre géométrie a procédé par hasard ». Il veut « donner aux conceptions géométriques cette extension et cette généralité qui sont dans sa nature et les constituer en une doctrine organique ». Il souligne aussi l'avantage qu'offre la géométrie projective du fait que ((jamais on n'y tire des conséquences sans que les formes réelles et existantes ne puissent se peindre à l'image et à la vue ».

La géométrie projective de Poncelet se fonde sur deux principes fondamentaux, les principes de continuité et de projection.

Le principe de continuité s'énonce ainsi : Chaque fois qu'une démonstration a été obtenue en supposant finies et réelles certaines parties de la figure qui interviennent dans la démonstration, la proposition subsiste quand ces parties disparaissent ou deviennent infinies ou imaginaires ou que la démonstration ne subsiste plus. Ainsi, l'extension du fini à l'infini réalisée par Desargues et Pascal est complétée par l'extension aux points imaginaires que ceux-ci n'avaient aucunement envisagée. Ce principe devait être vivement critiqué, particulièrement par Augustin Cauchy. Il convenait de préciser, ce qui fut fait plus tard, que ce principe vaut seulement lorsque les propriétés envisagées se traduisent par des relations algébriques, qui étaient d'ailleurs les seules que faisait intervenir Poncelet. L'application la plus obvie de ce principe est offerte par les intersections de deux cercles dont l'axe radical, défini par

les deux points d'intersection des cercles lorsque ceux-ci se coupent, subsiste encore lorsqu'ils ne se coupent plus.

Pour le principe de projection, Poncelet retient les propriétés qui se conservent par projection centrale et par les transformations qui en dérivent, notamment l'homologie dans le plan définie plus haut ; ces propriétés peuvent ètre réduites par projection « à des circonstances plus simples » se trouvant alors plus aisément démontrables. Michel Chasles (1793-l 880) devait exprimer ce principe d'une manière plus claire et le généraliser : « Que l'on prenne une figure quelconque de l'espace et l'une de ses propriétés communes, qu'on applique à l'une de ces figures l'un de ces modes de transformations et qu'on suive les diverses modifications qu'éprouve le théorème qui exprime cette propriété, on aura une nouvelle figure et une nouvelle propriété qui correspondra à celle de la première. Ce moyen que possède la géométrie récente permet de multiplier à l'infini les propriétés géométriques. »

La mise en œuvre de ces deux principes par Poncelet et, après lui, par plusieurs autres géomètres, Chasles en particulier, devait permettre une unification et une extension remarquables de la géométrie, spécialement dans le domaine de la théorie des coniques. Poncelet utilisa surtout la transformation qualifiée par lui d'homologie (définie plus haut), dont le principe avait été posé, on l'a vu, par Desargues, mais qu'il devait étendre à l'espace.

Chasles, posant le problème plus large de la détermination de la transformation ponctuelle la plus générale qui fait correspondre à une droite une droite et à un plan un plan, devait définir la transformation qu'il désigna par le terme d'homographie. K. G. C. von Staudt (1798-1867) et Gaston Darboux (1842-19 17) en donnèrent plus tard une justification plus rigoureuse. Cette transformation se définit analytiquement dans le plan par l'équation :

$$x' = \frac{ax + by + c}{dx + ey + f}$$

et par une équation analogue pour y'. L'homologie correspond au cas particulier où une droite l'axe d'homologie se transforme en elle-même, point par point.

Il faut en outre mentionner l'introduction par Monge de la transformation par polaire réciproque, qui, une conique (resp. une quadrique) étant donnée, fait correspondre réciproquement dans le plan un point (pôle) à une droite (sa polaire) et, dans l'espace, un point pôle à un plan (son plan polaire). Cette transformation fut systématiquement utilisée, notamment par Joseph Diez Gcrgonnc (1771-I 859), Brianchon, Poncelet et Chasles, qui devaient en montrer toute la fécondité. Par là était introduite la notion de dualité dont l'importance majeure ne tarda pas à être soulignée, et qui devait recevoir une extension bien au-delà de la géométrie projective, par l'introduction, due à Monge, mais non développée par lui, de la notion de transformation de contact.

En dépit des vues profondes de Poncelet et de Chasles et des nombreux résultats auxquels ils parvinrent, la géométrie projective souffrait d'un grave défaut : la distinction entre propriétés métriques et propriétés projectives n'était pas élucidée de façon satisfaisante, les propriétés projectives demeurant d'ailleurs, le plus souvent, définies par des considérations et des relations de caractère métrique. Cette carence devait conduire von Staudt, en 1847, à une présentation abstraite de la géométrie projective où n'intervenaient plus les éléments de caractère métrique, c'est-à-dire les notions d'angle et de dis-

tance. Cette synthèse n'était cependant pas tout à fait sans défaut. En particulier, elle faisait intervenir inutilement le postulat des parallèles. C'est seulement avec les travaux de Hilbert, Klein (1849-1925), Darboux (1842-1917) à la fin du XIXe siècle, et, plus tard, ceux de Federigo Enriques (1871-1946), que seront formulées de façon rigoureuse les notions de base et les axiomes de la géométrie projective.

4. Les géométries non euclidiennes

Jusqu'au début du XVIIIe siècle, le problème posé par le postulat des parallèles fut envisagé dans la même perspective : le postulat n'est pas une évidence première, mais une vérité qu'on doit pouvoir démontrer. La plupart des démonstrations se fondent sur la définition de la parallèle comme droite équidistante à une droite donnée, définition que l'on ne trouve pas dans les ***Éléments*** d'Euclide, il faut le noter. On ne soupçonne pas le cercle vicieux qu'implique une telle façon de faire, la possibilité qu'une droite puisse être équidistante à une autre droite supposant le postulat. Telles se présentent les démonstrations de Posidonius (IIe siècle av. J.-C.), de Geminus (I^e siècle apr. J.-C.), de Proclus (Ve siècle apr. J.-C.) ; et encore celle du jésuite Clavius (1537-1612) à la fin du XVIe siècle, celui-ci doutant cependant de la validité de sa démarche. Le jésuite G. Saccheri, dans son *Euclides ah omni naevo vindicatus* (1733), est le premier mathématicien à mettre nettement en doute la validité des démonstrations fondées sur l'équidistance et à proposer une autre approche, la réduction à l'absurde : supposer que le postulat des parallèles ne vaut pas et démontrer que cette hypothèse aboutit à une contradiction. À cet effet, Saccheri fait appel au trapèze isocèle qu'avait introduit le géomètre arabe Nāṣir al-Dīn (XIIIe siècle). Ce trapèze est construit en menant perpendiculairement aux extrémités d'une droite AB deux segments égaux AC et AD. Les angles intérieurs en C et D sont égaux. Ils valent un droit dans le cas où le postulat est vrai ; dans le cas contraire, ils sont soit aigus, soit obtus. Ainsi, bien avant que ne soit prouvée la validité des géométries non euclidiennes, était mis en évidence le dédoublement de l'hypothèse de la négation du postulat : angle aigu (géométrie de Lobatchevski), angle obtus (géométrie de Riemann). Nāṣir al-Din avait très vite cru pouvoir conclure que ces angles intérieurs en C et D étaient droits, pensant donc avoir démontré le postulat. Saccheri arrive à la même conclusion, mais au terme de longs développements, au cours desquels, contre son gré pourrait-on dire, il édifie pour une grande part la géométrie de Nikolaï Ivanovitch Lobatchevski (1792-1856) et pour une moindre part celle de Bernhard Riemann (1826-1866). Finalement, il rejette les deux hypothèses de l'angle aigu et de l'angle obtus, car elles le conduisent à deux conclusions qu'il estime non admissibles, la première à l'existence d'une perpendiculaire commune à deux droites à l'infini, la seconde à l'affirmation que deux droites contiennent un espace.

Plus de trente ans plus tard, en 1766, le mathématicien suisse Johann Heinrich Lambert (1728-1777), indépendamment semblet-il de Saccheri, dans une étude qui ne sera publiée qu'en 1786, suit fondamentalement la même démarche que Saccheri. Mais si, comme ce dernier, il rejette l'hypothèse de l'angle obtus, il est plus hésitant dans le cas de l'angle aigu.

Carl Friedrich Gauss (1777-1855) amorce vraiment, autour des années 1820,

la rupture avec la croyance bimillénaire en la démonstrabilité du postulat des parallèles : Gauss pense que l'on peut démontrer de façon rigoureuse que l'hypothèse de l'angle obtus conduit à une contradiction, mais il arrive à la conviction que l'on ne peut pas y parvenir dans le cas de l'hypothèse de l'angle aigu. Cette vue entraîne un changement radical dans la conception de la géométrie. Gauss déclare que, désormais, « la géométrie ne doit pas être mise au même rang que l'arithmétique dont la vérité est purement a priori, mais plutôt au même rang que la mécanique ».

Lobatchevski, dont les travaux se situent entre 1826 et 1856, parvient en 1834 à une conclusion encore plus explicite : « La vérité à établir le postulat des parallèles n'est pas impliquée dans les notions antérieures ; pour la démontrer, il faut recourir à des expériences, par exemple aux observations astronomiques. » Ces expériences furent faites à l'époque et montrèrent qu'au degré de précision des appareils de mesure on ne pouvait écarter la géométrie euclidienne. Le mathématicien hongrois Farkas Bolyai (1775-1856), qui, indépendamment de Lobatchevski. élabore de façon assez complète la géométrie de l'angle aigu dans *La Science absolue de l'espace* (1832), n'a pas une attitude aussi nette ; il n'aperçoit pas clairement que la validité du postulat des parallèles est à chercher non dans une déduction logique à partir des axiomes d'Euclide, mais dans l'expérience.

Quant à l'hypothèse de l'angle obtus, en 1854, elle est reconnue acceptable par Riemann, bien qu'elle conduise à affirmer que les droites sont de longueur finie et que deux droites peuvent enfermer un espace.

Toutefois, ces vues nouvelles ne firent qu'assez lentement leur chemin. Il restait d'ailleurs à s'assurer qu'en poursuivant le développement des deux géométries non euclidiennes, on n'y rencontrerait pas de contradiction, ce qui ne fut réalisé de façon pleinement satisfaisante qu'à la fin du XIXe siècle grâce aux travaux de Klein et par l'élaboration de modèles des deux géométries ; ainsi, pour la géométrie de Bolyai-Lobatchevski, par le modèle de Henri Poincaré (1854-1912), où l'on considère le demi-plan et où les droites sont représentées par les demi-cercles centrés sur la droite qui limite ce demi-plan ; et, pour la géométrie de Riemann, par une correspondance associant à un point de l'espace une droite, et à une droite un plan (cf. **GROUPES** - Groupes classiques et géométrie, chap. 3).

Soit qu'au début ils les rejettent comme aboutissant à des contradictions, soit que, au contraire, ne parvenant pas à y trouver de contradiction, ils inclinent à les reconnaître valables, durant un siècle environ, à partir de Saccheri jusqu'à Lobatchevski et Bolyai, les géomètres ont peu à peu élaboré la géométrie de l'angle aigu et, de façon beaucoup plus sommaire, celle de l'angle obtus qui ne prend vraiment consistance qu'après son acceptation par Riemann.

Saccheri et Lambert montrent que les trois hypothèses, angle aigu, angle droit. angle obtus, sont stables, c'est-à-dire que, si elles valent pour un trapèze, elles valent pour tout trapèze. Ils établissent en outre que ces trois hypothèses sont équivalentes aux trois catégories de valeurs possibles pour la somme des angles d'un triangle : inférieure, égale, supérieure à deux droits.

Dans le cas de l'angle aigu, Saccheri montre, ce qu'avait à peine aperçu Lambert, que, pour un point situé hors d'une droite, on peut mener une infinité de droites non sécantes. Pour chaque droite non sécante, on peut déterminer une perpendiculaire commune à cette non-

sécante et à la droite. De plus, Saccheri a reconnu l'existence de deux droites limites qui se rapprochent indéfiniment de la droite donnée, l'une à droite, l'autre à gauche.

Allant plus loin que Saccheri, Lambert, se fondant sur l'analogie de la géométrie de l'angle aigu avec la géométrie sur la sphère, pour la valeur de la surface d'un triangle, montre, en faisant appel à une sphère imaginaire, que, dans la géométrie de l'angle aigu, la surface d'un triangle est proportionnelle à la différence entre deux droits et la somme des angles du triangle. Par là, il introduit une constante caractéristique de l'espace qui séduira Gauss et constituera une des raisons principales qui inciteront le mathématicien allemand à penser que cette géométrie peut être vraie. Lambert est conduit à une conclusion assez surprenante, mais qui ne le choque pas : dans la géométrie de l'angle aigu, quand la longueur des trois côtés d'un triangle devient infinie, la surface du triangle n'en demeure pas moins finie.

Quant à la géométrie de l'angle obtus, dont Riemann montre en 1854 qu'elle n'est nullement contradictoire, un esprit moderne peut s'étonner qu'il ait fallu attendre si longtemps avant de l'admettre, alors que, dès le milieu du XVIIIe siècle, les études sur les propriétés infinitésimales des surfaces et même, bien avant, la connaissance des triangles sphériques en offraient un modèle dans le cas de l'espace à deux dimensions, au moins dans le voisinage d'un point. Mais une telle manière de voir les choses est anachronique : elle implique qu'une surface peut être considérée comme constituant un espace à deux dimensions. C'est là une vue très moderne. Encore au XVIIIe siècle et même au XIXe, une surface ne pouvait aucunement être considérée comme constituant un espace ; elle n'était qu'une figure de l'espace dont on ne concevait pas qu'il pût avoir une structure ; on le regardait seulement comme un cadre homogène et infini.

5. Transformations géométriques

En introduisant la projection centrale, ou perspective, en géométrie, Desargues puis Pascal avaient ouvert la voie à l'étude des transformations géométriques. Ce n'est qu'à la fin du XVIIIe siècle que les transformations géométriques commencèrent vraiment à retenir l'attention des mathématiciens. À côté de la projection centrale, l'homologie est systématiquement utilisée comme transformation par Poncelet ; puis Chasles définit l'homographie, transformation projective la plus générale. D'autre part, des transformations plus particulières sont largement étudiées : l'affinité, à laquelle s'était déjà intéressé Euler, les rotations, les symétries, les translations, les homothéties.

Chez les géomètres purs, les transformations apparaissent surtout comme un instrument de démonstration, tout spécialement chez Chasles. Mais les mathématiciens qui, comme Arthur Cayley (1821-1895) notamment, s'intéressent surtout à l'analyse et à l'algèbre, s'attachent à leurs aspects d'invariance. Ainsi s'amorce entre l'algèbre et la géométrie une symbiose féconde.

Ces divers types de recherches tendent à donner aux transformations une place non plus marginale, mais centrale en géométrie, au point que l'on voit au milieu du XIXe siècle se dégager l'idée que les propriétés géométriques se classent et se caractérisent par les transformations qui les laissent invariantes. À chaque type de transformation correspond une géométrie.

Ainsi, en 1868, Hermann von Helmoltz développe l'idée que l'on peut caractériser les propriétés de l'espace euclidien par les propriétés des déplacements, envisagés comme transformations ponctuelles. D'autre part, la notion de transformation allait permettre de préciser les relations entre propriétés projectives et propriétés métriques (que ni Poncelet, ni Chasles, ni même von Staudt n'avaient vraiment élucidées) et, en outre, entre géométrie euclidienne et géométrie non euclidienne. Le premier pas dans cette voie est fait par Cayley, en 1859 : ayant particularisé la transformation homographique en lui imposant la conservation d'une conique dans le plan ou d'une quadrique dans l'espace, il peut définir la distance de deux points A et B comme le logarithme du rapport anharmonique de ces deux points et des points de rencontre de la droite qui les porte avec la conique ou la quadrique. Cette approche allait se révéler très féconde. Edmond Laguerre (1834-1 886) s'était déjà engagé dans cette voie en 1853 lorsqu'il avait lié la mesure d'un angle au rapport anharmonique de ses côtés et des deux droites isotropes passant par le point de rencontre de ces côtés.

Cayley montre alors que l'on peut déterminer une conique ou une quadrique telle que la distance projective correspondant à la transformation projective qui la conserve est identique à la distance de la géométrie euclidienne. La géométrie euclidienne apparaît comme un type particulier de géométrie projective. Enthousiasmé par ce résultat, Cayley déclare à la fin de son mémoire : « La géométrie projective est toute la géométrie. »

Dans des travaux entrepris à partir de 1868, Klein reprend ces vues pour les préciser et leur donner de plus larges applications. Il définit de façon plus rigoureuse la notion de distance cayleyenne ; puis, portant son attention aux deux géométries non euclidiennes, dont ne s'était pas occupé Cayley, il montre qu'elles viennent aussi prendre place dans le cadre de la géométrie projective, en associant à chacune des deux géométries non euclidiennes une transformation conservant une conique dans l'espace à deux dimensions, et une quadrique dans l'espace à trois dimensions, définies de façon appropriée. Ainsi les trois géométries d'Euclide, de Bolyai-Lobatchevski et de Riemann se trouvent rapprochées et apparaissent comme des types particuliers de géométrie projective.

De plus, posant le problème général de la détermination des géométries projectives à courbure constante, Klein établit qu'il ne peut en exister que trois types qui sont précisément les trois géométries. Ainsi les géométries non euclidiennes, qui étaient considérées jusque-là comme des situations géométriques plutôt bizarres, aberrantes et sans rapport avec la géométrie euclidienne et la géométrie projective, se révèlent avoir une place précise et nécessaire au sein de la géométrie projective à côté de la géométrie euclidienne ; Klein tient à noter que cette conclusion est obtenue par des considérations « tout autres » que celles par lesquelles ces géométries avaient été précédemment définies. Ajoutons que, par une telle intégration, il était clairement établi que la géométrie projective générale ne fait pas intervenir le postulat des parallèles, ce que ni Poncelet, ni Chasles, ni Staudt n'avaient su montrer.

Transformations et groupes

Le rôle des transformations, en géométrie, ne fut pleinement compris que lorsque Klein leur associa la notion de groupe,

introduite par Évariste Galois (18 1 l-l 832) en 1830, et diffusée seulement en 1870 par le *Traité des substitutions et des équations algébriques* de Camille Jordan (1838-1922). C'est par cet ouvrage que Klein en prit connaissance. Certes, en 1844, Hermann Grassmann (1809-l 877) avait déjà pressenti qu'il y aurait lieu de faire intervenir l'algèbre dans l'étude fondamentale de la géométrie, mais il ne devait pas dépasser dans cette voie des vues générales. Klein, au contraire, envisage d'emblée une « théorie des groupes qui peuvent être engendrés par les transformations d'une nature donnée ». Il montre alors que la plupart des transformations géométriques considérées avant lui constituent bien des groupes (loi de composition, élément inverse, élément unité). Ces groupes sont hiérarchisés. Klein appelle groupe principal celui qui correspond aux transformations qui n'altèrent pas les propriétés géométriques des figures. Il est défini par l'ensemble des opérations de translation, de rotation et de symétrie. En supprimant les symétries, on obtient le groupe des déplacements euclidiens. Le groupe principal lui-même s'insère dans le groupe des similitudes constitué par l'ensemble des opérations d'homothéties et de translation. Ce groupe prend place à son tour dans le groupe affine qui garde le parallélisme et, pour une direction donnée, les rapports de longueur. On peut le définir comme une homographie conservant le plan à l'infini. Ainsi, alors que demeuraient auparavant nettement distinguées et considérées comme de nature différente les opérations qui n'altèrent pas les figures, qui préservent en quelque sorte leur identité, et, à l'opposé, des opérations qui les modifient, comme l'étaient déjà les symétries et les similitudes, mais beaucoup plus encore les homologies et les homographies, Klein propose, par une mutation de perspective dont on ne saurait trop souligner l'audace et la portée, de ne plus s'attacher à une telle distinction et de réunir toutes ces opérations dans un type unique de processus : des transformations hiérarchisées constituant des groupes. Cette conception systématique et unifiée est proposée par lui en 1872 dans le célèbre programme d'Erlangen. L'objet de la géométrie se trouve ainsi défini : « Étant donné une multiplicité et un groupe, étudier les êtres au point de vue des propriétés qui ne sont pas altérées par les transformations du groupe », ou encore : « Étant donné une multiplicité et un groupe de transformation, développer la théorie des invariants relatifs à ce groupe. » Comme le dit encore Klein, « les méthodes géométriques modernes sont caractérisées par le fait que leurs considérations, au lieu de s'appuyer sur le groupe principal, reposent sur des groupes de transformation plus étendus. Dès que leurs groupes se contiennent l'un l'autre, une loi analogue établit leurs rapports réciproques. De la sorte, pour la première fois, les divers ordres de recherche de la géométrie sont exprimés par des groupes de transformation qui leur correspondent. »

Cette doctrine unificatrice ne se limitera pas aux transformations que l'on avait étudiées jusque-là, et dont la plus générale était l'homographie. La logique même de la démarche de Klein et ses contacts avec le mathématicien norvégien Sophus Lie (1842-I 899), qui venait d'entreprendre ses travaux sur les groupes continus, conduisaient à envisager des groupes plus généraux allant jusqu'au groupe le plus général des transformations continues auquel correspondent des propriétés de « position », dont plusieurs avaient déjà été reconnues et dont l'étude constituera la topologie algébrique.

6. La généralisation de Riemann

Si la synthèse de Klein avait pu sembler couvrir tout le champ de la géométrie, elle ne portait en fait que sur des espaces « homogènes ». En 1854, dans sa *Dissertation inaugurale*, Riemann avait proposé une conception plus générale, et, en un sens, plus profonde de la géométrie. Il s'était attaché au « concept général de grandeur de dimension multiple (déjà envisagé, mais de façon beaucoup moins précise, par Grassmann en 1847), comprenant comme cas particulier les grandeurs étendues » et il avait conclu qu'« une grandeur de dimensions multiples est susceptible de différents rapports métriques » et que « l'espace n'est par suite qu'un cas particulier d'une grandeur de trois dimensions ». Se libérant encore plus ncttcmcnt que Gauss et Lobatchevski de la limitation qu'imposait le lien de la géométrie avec l'espace physique, et allant au-delà de la reconnaissance de la validité de la géométrie de l'angle obtus, Riemann fut amené à définir un espace à partir d'éléments différentiels, en exprimant le carré ds^2 de l'élément de distance entre deux points infiniment voisins en fonction des éléments différentiels dx, dy et dz des coordonnées d'un point. Riemann définissait alors un type d'espace très général dont l'intérêt devait notamment apparaître lors de la création par Albert Einstein de la théorie de la relativité générale. Les conceptions de Klein et de Riemann ne furent « raccordées » que par les travaux d'Élie Cartan (1869-1951), qui généralisa la notion d'espace de Riemann par l'introduction d'une connexion définie par un groupe.

Déjà avec Klein, mais plus encore avec Riemann, la notion de géométrie avait connu une mutation profonde, non seulement par les généralisations qu'ils en donnaient, mais aussi par la substitution, dans la conception de la géométrie, à l'intérêt premier porté aux figures, de la considération de la nature même de l'espace. Réalité « neutre », sans « forme », simple réceptacle dans la géométrie classique, l'espace, désormais envisagé comme une structure, était reconnu comme le constituant fondamental de la géométrie, comme son objet premier.

FRANÇOIS RUSSO

Bibliographie

M. BERGER, *Géométrie*, 2 vol., Nathan, Paris, 1990 / M. CHASLES, *Aperçu historique sur l'origine et le développement des méthodes en géométrie*, Gauthier-Villars, 1889, repr. en fac-sim., Gabay, Paris, 1989 / G. CHOQUET, *L'Enseignement de la géométrie*, Hermann, Paris, 1964 / H. S. M. COXETER, *Introduction to Geometry*, John Wiley, New York. 2e éd. 1989 / *Encyclopédie des sciences mathématiques pures et appliquées*, t. III, 2 vol., Gauthier-Villars et Teubner. Paris, Leipzig. 1911, repr. en fac-sim. Gabay, 1992 / J. GRAY, *Ideas of Space*, Clarendon Press, Oxford. 2e éd. 1989 / D. HILBERT, *Les Fondements de la géométrie*, éd. critique par Paul Rossier, Dunod. Paris, 1971 / D. HILBERT & COHN-VOSSEN, *Geometry and Imagination*, Chelsea, New York, 1952 / F. KLEIN, *Le Programme d'Erlangen*, Gauthier-Villars, Paris. 1974, repr. en fac-sim. Gabay, 1991 / M. KLINE, *Mathematical Thought from Ancient to Modem Times*, 3 vol., Oxford Univ. Press, New York, 1990 / J. LELONG-FERRAND, *Les Fondements de la géométrie*, P.U.F.. Paris. 1985 / R. TATON, *L'Œuvre mathématique de Desargues*, J. Vrin. Paris, 1988 ; *L'Œuvre scientifique de Monge*, Paris, 1951 / R. TATON dir., *Histoire générale des sciences*, 3 vol., P.U.F., 2e éd. 1969-1983 / C. TISSERON. *Géométries affine, projective et euclidienne*, Hermann, Paris. 1983.

GÉOMÉTRIE ALGÉBRIQUE

Sous sa forme actuelle, la géométrie algébrique est une branche de l'algèbre relativement récente (cf. ALGÈBRE). Pour

« comprendre » les phénomènes d'intersection des courbes et des surfaces, il s'est révélé nécessaire d'élaborer des techniques compliquées qui se sont développées de manière abstraite et sont venues à leur tour enrichir d'autres domaines des mathématiques (théorie moderne des nombres, fonctions analytiques de plusieurs variables complexes, topologie algébrique) ; pour le profane, cet appareil mathématique peut sembler bien loin de l'« intuition géométrique » !

La géométrie algébrique est issue de l'étude des courbes algébriques du plan $\mathbf{R}^2$ ou de l'espace $\mathbf{R}^3$ et des surfaces algébriques de $\mathbf{R}^3$. Pendant le XVIIIe et le XIXe siècle, on s'est aperçu qu'il était plus commode de modifier le problème en se plaçant dans le plan complexe $\mathbf{C}^2$ ou dans l'espace complexe $\mathbf{C}^3$; en effet, C est un corps algébriquement clos, de sorte que les courbes et les surfaces ont toujours « suffisamment » de points à coordonnées complexes, alors qu'il peut n'y avoir aucun point à coordonnées réelles (comme c'est le cas pour la courbe d'équation $x^2 + y^2 + 1 = 0$). On a observé aussi que certains énoncés intéressants n'étaient vrais que si l'on complétait les courbes et les surfaces par des « points à l'infini », se plaçant ainsi dans le plan projectif $\mathbf{P}_2(\mathbf{C})$ ou dans l'espace projectif $\mathbf{P}_3(\mathbf{C})$; les courbes ou les surfaces y sont définies par des équations polynomiales homogènes portant sur les coordonnées homogènes.

Cette diversité de points de vue (réel ou complexe, affine ou projectif) a dû être encore élargie lorsque la théorie des nombres a mis en évidence l'intérêt de l'étude des courbes algébriques définies sur des corps autres que R ou C, comme les corps finis ou les corps p-adiques ; la théorie des équations diophantiennes conduit même à considérer des courbes ou des ensembles algébriques définis sur un anneau tel que Z. Pendant la première moitié du XXe siècle, l'école allemande a développé la théorie des ensembles algébriques (de dimension quelconque) de l'espace affine k^n ou de l'espace projectif P,(k), k étant un corps de base algébriquement clos arbitraire.

Pour l'étude des propriétés intrinsèques d'un ensemble algébrique, il est plutôt gênant d'avoir à le considérer comme plongé dans un espace affine ou un espace projectif. Le problème se pose donc de définir des « variétés algébriques abstraites », non plongées dans k^n ou P,(k), un peu comme on définit des variétés différentiables indépendamment d'un plongement dans $\mathbf{R}^n$ en géométrie différentielle. De telles variétés abstraites ont été définies par A. Weil (1946). Une définition équivalente, plus simple et plus maniable se trouve dans l'article de J.-P. Serre, « Faisceaux algébriques cohérents » (1955) ; elle est inspirée de la théorie des espaces analytiques. Dans le présent article, nous donnerons la définition de Serre un peu élargie, en prenant comme corps de base un corps algébriquement clos. Le cas d'un corps de base non algébriquement clos, ou d'une base plus générale, s'exprime bien dans le cadre de la théorie des schémas de A. Grothendieck, qui généralise considérablement celle des variétés algébriques au sens de Serre en partant du même point de vue.



1. Ensembles algébriques

Soit k un corps de base algébriquement clos. Pour tout entier naturel n, l'espace affine k^n est l'ensemble des suites $(x_1, x_2, \ldots, x_n)$ de n éléments de k ; on appelle ces n

éléments les *coordonnées* du point $s = (x_1, x_2, \ldots, x_n)$ de k^n. L'espace projectif $P_n(k)$ est le quotient de $k^{n+1} - \{O\}$, complémentaire de l'origine $0 = (0, 0, \ldots, 0)$, par la relation d'équivalence qui identifie $(x_0, x_1, \ldots, x_n)$ à toute suite proportionnelle $(tx_0, tx_1, \ldots, tx_n)$ (t élément non nul de k) ; on voit que les points $P_n(k)$ sont les droites passant par 0 dans k^{n+1}, privées de 0. Si un élément $(x_0, x_1, \ldots, x_n)$ de k^{n+1} représente un point x de $P_n(k)$, on dit que les coordonnées $x_0, x_1, \ldots, x_n$ de cet élément sont des *coordonnées homogènes* de x ; elles ne sont pas toutes nulles et sont définies à un facteur de proportionnalité près.

Une partie X de k^n est un *ensemble algébrique affine* si c'est l'ensemble des zéros communs à des polynômes $f_1, f_2, \ldots, f_s$ par rapport aux coordonnées ; on dit que X est défini par les équations :

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \ldots\ldots\ldots\ldots\ldots \\ f_s(x_1, x_2, \ldots, x_n) = 0. \end{cases}$$

Un *ensemble algébrique projectif* dans $P_n(k)$ est défini d'une manière analogue par des équations polynomiales homogènes par rapport aux coordonnées homogènes.

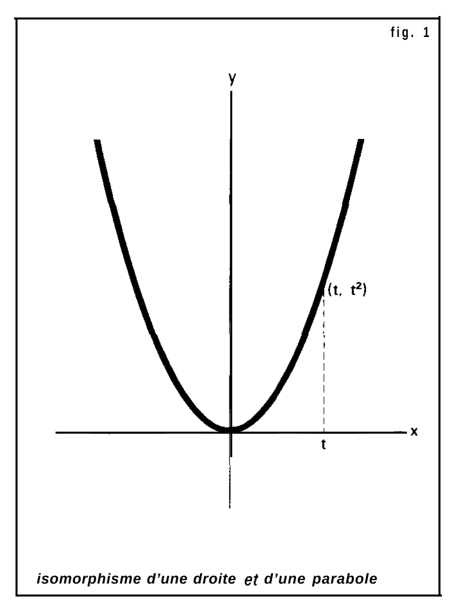
Applications régulières

Soient $X \subset k^m$ et $Y \subset k^n$ des ensembles algébriques affines ; une application u de X dans Y est dite *régulière* si les coordonnées $u_1(x), u_2(x), \ldots, u_n(x)$ de $u(x)$ sont des fonctions polynomiales des coordonnées du point x de X. En particulier, les applications régulières de X dans k, encore appelées *fonctions régulières* sur X, sont les fonctions polynomiales des coordonnées d'un point de X ; elles forment un sous-anneau de l'anneau de toutes les applications de X dans k, et ce sous-anneau est visiblement isomorphe au quotient de l'anneau des polynômes $k[T_1, T_2, \ldots, T_m]$ par l'idéal des polynômes qui s'annulent sur X.

Il est clair que la composée de deux applications régulières est une application régulière. En particulier, une application régulière $u : X \rightarrow Y$ définit un homomorphisme $f \mapsto f \circ u$, de l'anneau des fonctions régulières sur Y dans l'anneau des fonctions régulières sur X. Un *isomorphisme* d'un ensemble algébrique X sur un autre Y est une application bijective de X sur Y, qui est régulière ainsi que sa réciproque ; il définit un isomorphisme de l'anneau des fonctions régulières sur Y sur l'anneau des fonctions régulières sur X.

On trouvera ci-après quelques exemples d'applications régulières.

Paramétrisation d'une parabole (fig. 1). Considérons l'application u de la droite k

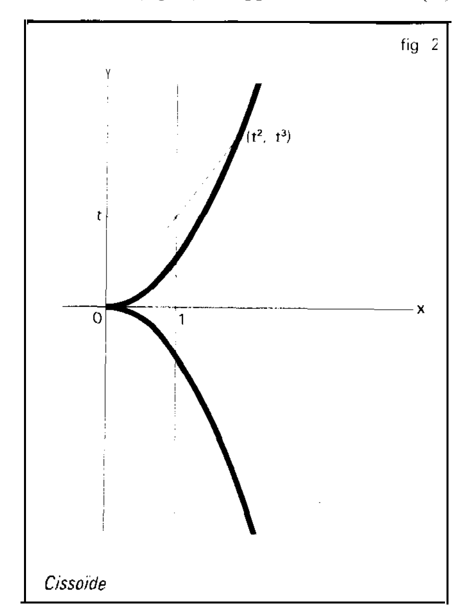


isomorphisme d'une droite et d'une parabole

dans le plan k^2 définie par $u(t) = (t, t^2)$. L'image $u(k)$ est la parabole X d'équation $y = x^2$, et u définit une bijection de k sur

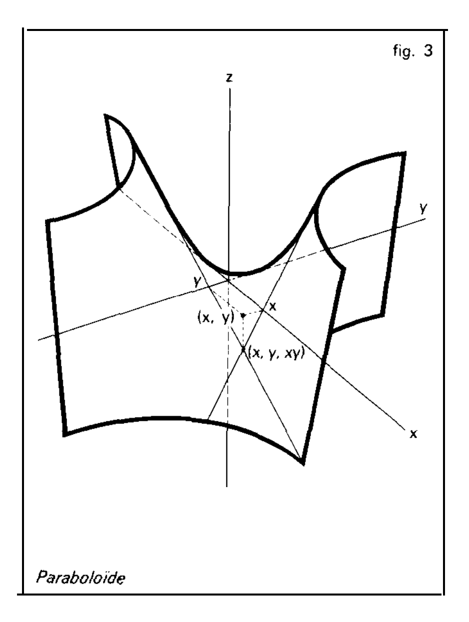
X, réciproque de l'application $(x, y) \mapsto x$; on a donc un isomorphisme de la droite k sur la parabole X.

– *Cissoïde* (fig. 2). L'application $v : t \mapsto (t^2, t^3)$ de k dans k^2 est aussi une application régulière. Elle applique la droite k bijectivement sur son image, qui est la « parabole semi-cubique » (ou cissoïde) Y d'équation $y^2 = x^3$; mais ce n'est pas un isomorphisme de k sur Y, car la bijection réciproque v', définie par $v'(x, y) = y/x$ si $x \neq 0$ et $v'(0, 0) = 0$ n'est pas une application régulière. En fait on peut montrer que la cissoïde Y n'est pas isomorphe à une droite en observant que son anneau de fonctions régulières n'est pas intégralement clos (la fraction $z = y/x$ vérifie l'équation $z^2 = x$, donc est entière sur cet anneau, sans y appartenir, cf. ANNEAUX COMMUTATIFS), et ne peut par suite être isomorphe à l'anneau k[T] des fonctions régulières sur k.



Cissoïde

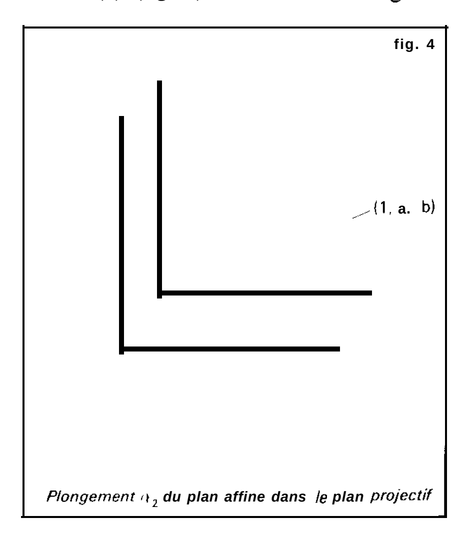
– *Paraboloïde* (fig. 3). L'application rationnelle (x, y) $\mapsto (x, y, xy)$ du plan k^2 dans l'espace k^3 définit un isomorphisme du plan sur le paraboloïde d'équation $z = xy$.



Paraboloïde

Pour définir les applications régulières d'un ensemble algébrique projectif $X \subset \mathbf{P}_m(k)$ dans un autre $Y \subset \mathbf{P}_n(k)$, on procède de la même façon. Une application u de X dans Y est régulière si les coordonnées homogènes de $u(x)$ peuvent s'exprimer par des polynômes homogènes $u_0, u_1, ..., u_n$, tous de même degré par rapport aux coordonnées homogènes de x ; notons que ces polynômes ne peuvent s'annuler simultanément pour $x \in X$. Si maintenant X C k^m est affine et Y C P,(k) est projectif, les applications régulières de X dans Y sont définies en donnant les coordonnées homogènes sur Y comme fonctions polynomiales (ne s'annulant pas simultanément) des coordonnées sur X ; quant aux applications régulières de Y dans X, ce sont les applications constantes (les polynômes homogènes de degré 0 sont les seuls à être constants sur chaque droite issue de l'origine dans k^{n+1}).

Par exemple, l'application α_n de k^n dans P,,(k) qui transforme $(x_1, x_2, \ldots$ x,) en le point de coordonnées homogènes $(1, x_1, \ldots x_n)$ est régulière ; elle définit une bijection de k^n sur le complémentaire de I'hyperplan d'équation homogène $x_0 = 0$ dans P,(k) (fig. 4). Tout ensemble algébri-



Plongement α_2 du plan affine dans le plan projectif

que X de k^n s'identifie par α_n à la partie d'un ensemble algébrique $\overline{X}$ de P,(k) où $x_0 \neq 0$; si les équations de $\overline{X}$ sont :

$$f_1 = 0, \ldots, f_s = 0$$

les équations homogènes de $\overline{X}$ sont :

$$x_0^{d_1} f_1(x_1/x_0, \ldots, x_n/x_0) = 0, \ldots x_0^{d_s} f_s(x_1/x_0, \ldots x_n/x_0) = 0$$

(d_i désigne le degré f_i) ; nous dirons que $\overline{X}$ est la complétion projective de X. Toute application régulière u d'un ensemble algébrique affine X dans un autre Y se prolonge d'une manière unique en une application régulière $\bar{u}$ de $\overline{X}$ dans $\overline{Y}$. En particulier, l'application $u : t \mapsto (t, t^2)$ de k dans k^2 se prolonge en l'application régulière $\bar{u}$: P,(k) $\mapsto$ *P,(k),* qui transforme le point de coordonnées homogènes $(x$, y) en le point de coordonnées homogènes (x^2, xy, y^2) ; cette application définit un isomorphisme de la droite projective *P,(k)* sur la conique $\overline{X}$ complétée de la parabole X (cf. exemple ci-dessus), dont l'équation homogène est y^2 = xz.

On peut généraliser l'exemple précédent et l'on peut ainsi définir pour tout couple (n, d) d'entiers naturels une application régulière de P,(k) dans P,(k) dans laquelle $e = (n + 1)(n + 2)\ldots(n + d)/$ d ! $-$ 1, de manière à obtenir un isomorphisme de P,(k) sur un ensemble algébrique $V_{n,d}$ dans P,(k) (la variété de Veronese) ; les coordonnées homogènes du transformé de x sont tous les monômes de degré *d* par rapport aux coordonnées homogènes de x.

Applications rationnelles

En remplaçant les polynômes par des fractions rationnelles dans tout ce qui précède, on obtient des « applications » non partout définies en général (car les dénominateurs peuvent s'annuler ; il y a donc un abus de langage à parler d'applications) ; ce sont les applications rationnelles. Nous ne donnons de définition précise que dans le cas des *fonctions rationnelles* (applications rationnelles à valeur dans ***k*).** Considérons d'abord un ensemble algébrique affine X dans k^n ; une fonction rationnellefsur X est définie par une fraction $P/Q \in k(T_1, T_2, \ldots T_n)$, dont le dénominateur ne s'annule pas identiquement sur X ; c'est l'application $x \mapsto P(x)/Q(x)$ de l'ensemble $U = \{x \in X \mid Q(x) \neq 0\}$ dans k. De même, pour définir une fonction rationnelle sur un ensemble algébrique projectif XC P,,(k) on prend une fraction P/Q en coordonnées homogènes, avec P et Q homogènes *de même degré* (pour avoir une fonction constante sur les droites issues de

0 dans k^{n+1}) et Q non identiquement nul sur X.

La composée de deux applications rationnelles f de X dans Y et g de Y dans Z peut se définir si l'ensemble des points x de X tels quefsoit définie en x et g en $f(x)$ n'est contenu dans aucun ensemble algébrique strictement plus petit que X ; c'est encore une application rationnelle. Une *équivalence birationnelle* entre X et Y est un couple (u, v) où u est une application rationnelle de X dans Y et v une application rationnelle de Y dans X, les composés $v \circ u$ et $u \circ v$ étant les applications identiques de X et Y respectivement.

Par exemple $(x, y) \mapsto y/x$ est une fonction rationnelle dans le plan k^2, définie dans le complémentaire de la droite d'équation $x = 0$. Sa restriction à la courbe d'équation $y^2 = x^3$ est une fonction rationnelle définie en dehors de l'origine ; on voit que la cissoïde est birationnellement équivalente à la droite (sans lui être isomorphe ; cf. exemple supra et fig. 2). D'une manière générale, on dit qu'une courbe algébrique est unicursale si elle est birdtionnellement équivalente à la droite k (Cf. COURBES ALGÉBRIQUES). Les fractions $x_1/x_0, x_2/x_0, \ldots, x_n/x_0$ définissent une application rationnelle de $\mathbf{P}_n(k)$ dans k^n ; cette application est définie dans le complémentaire de l'hyperplan d'équation homogène $x_0 = 0$ et donne (avec α_n ; cf. *supra*) une équivalence birationnelle entre $\mathbf{P}_n(k)$ et k^n. De même, tout ensemble algébrique affine est birationnellement équivalent à sa complétion projective.

2. Variétés algébriques affines

À tout ensemble algébrique affine X C k^m, nous avons associé la k-algèbre A(X) des fonctions régulières sur **X** ; elle est isomorphe (d'une manière canonique) au quotient $k[T_1, T_2, \ldots T_m]/I(X)$ où I(X) désigne l'idéal formé des polynômes qui s'annulent sur X. Si une application $u : X \rightarrow Y$ d'un ensemble algébrique dans un autre est régulière, $f \circ u$ appartient à A(X) pour toute fonction f de A(Y). Inversement, cette condition implique que u est régulière ; remplaçons en effet f par les fonctions coordonnées $y_1, y_2, .., y_n$ de Y : nous obtenons des fonctions $u_i = y_i \circ u$ $(i = 1, 2, ..., n)$ régulières sur X, c'est-à-dire induites par des polynômes en les coordonnées de X.

On voit même que tout homomorphisme φ de A(Y) dans A(X) détermine une application régulière u de X dans Y telle que φ soit l'application $f \mapsto f \circ u$; les coordonnées de u sont les fonctions $\varphi(y_1), \varphi(y_2), \ldots \varphi(y_n)$ de A(X). Considérons, en particulier, le cas où X = $\{e\}$ est réduit à un point ; c'est l'espace affine k^0 et son algèbre de fonctions régulières se réduit aux constantes A(X) = k. La donnée d'une application (régulière automatiquement) u : x = $\{e\} \rightarrow$ Y, c'est-à-dire d'un point y = u(e) de Y, équivaut donc à celle de l'homomorphisme $f \mapsto f \circ u = f(y)$ de A(Y) dans k ; d'où une bijection de l'ensemble Y sur l'ensemble Horn, (A(Y), k) des homomorphismes de A(Y) dans k.

Tout isomorphisme A(Y) $\rightarrow$ A(X), où X et Y sont des ensembles algébriques affines, détermine un isomorphisme de X sur Y. Cela nous met sur la voie d'une définition intrinsèque des ensembles algébriques affines, indépendamment du plongement dans un espace k^n : la structure d'ensemble algébrique est définie par la donnée de l'algèbre des fonctions régulières. Nous allons considérer une structure un peu plus fine, en utilisant une autre algèbre qui n'est pas une algèbre de fonctions. En effet, il est avantageux de

pouvoir distinguer. par exemple, l'ensemble algébrique {O} C k défini par l'équation $x = 0$ (« point simple ») du même ensemble défini par l'équation $x^2 = 0$ (« **point double** »), bien que ces ensembles soient isomorphes. Pour cela, on est **conduit** à associer à l'ensemble algébrique X C k^n, défini par les équations $f_1 = 0$, $f_2 = 0$, ..., $f_s = 0$, non pas l'algèbre de fonctions $k[T_1, T_2, \ldots T_n]/I(X)$, mais l'algèbre $k[T_1, T_2, \ldots T_n]/\mathfrak{a}$, où a est l'idéal de polynômes engendré par $f_1, f_2, \ldots, f_s$; il est clair que a est contenu dans I(X), donc l'algèbre des fonctions régulières sur X s'identifie à un quotient de la nouvelle algèbre ; ainsi, tout élément de cette nouvelle algèbre définit une fonction régulière f sur X (sa classe modulo $I(X)/\mathfrak{a}$), mais f peut être nulle sans que l'élément considéré **le soit. Dans** l'exemple de {0} C k, l'algèbre associée est $k[T]/(T) \simeq k$ dans le cas du point simple, d'équation $x = 0$, et $k[T]/(T^2) \simeq k + k\varepsilon$, algèbre des *nombres duaux* (extension quadratique de k engendrée par un élément ε de carré nul) dans le cas du point double, d'équation $x^2 = 0$; l'élément ε définit une fonction nulle.

Nous appellerons *variété algébrique affine* un triplet (X, A, φ) où X est un ensemble, A une k-algèbre engendrée par un nombre fini d'éléments et φ une bijection de X sur $\mathrm{Hom}_k(A, k)$. Notons que si $(x_1, x_2, \ldots x_n)$ est un système de générateurs de A, il détermine un homomorphisme surjectif $k[T_1, T_2 \ldots T_n] \to A$ dont le noyau a est engendré par un nombre fini de polynômes, car l'anneau des polynômes est noethérien (théorème de Hilbert, cf. ANNEAUX COMMUTATIFS) ; ainsi A est isomorphe à l'algèbre associée à un ensemble algébrique affine contenue dans k^n, **et** φ détermine une bijection de X sur cet ensemble. Un *morphisme* (X, A, φ) $\to$ (Y, B, ψ) de variétés algébriques affines est un couple (u, v) d'une application $u : X \to Y$ et d'un homomorphisme $v : B \to A$ de k-algèbres, tel que $\varphi(x) \circ v = \psi(u(x))$ pour tout point x de X ; en fait la connaissance de v détermine entièrement u (à l'aide de φ et ψ). Le composé d'un tel morphisme avec un morphisme (u', v') : (Y, B, ψ) $\to$ (Z, C, χ) est le morphisme $(u' \circ u, v \circ v')$ de (X, A, φ) dans (Z, C, χ). Le morphisme (u, v) est un isomorphisme s'il existe un morphisme (u', v') de (Y, B, ψ) dans (X, A, φ) tel que les composés $(u', v') \circ (u, v)$ et $(u, v) \circ (u', v')$ soient les morphismes identiques (id_X, id_A) et (id_Y, id_B) ; cela revient à dire que v est un isomorphisme de k-algèbres.

La droite affine est la variété $(k, k[T], \psi)$ où ψ applique tout élément a de k sur l'homomorphisme $k[T] \to k$ qui transforme T en a. Un morphisme de (X, A, φ) dans la droite affine est donc formé d'une application u de X dans k et d'un homomorphisme v de k[T] dans A ; la donnée de v, qui équivaut à celle du morphisme (u, v), revient à celle de l'élément $f = v(T)$ de A ; pour tout point x de X on a $u(x) = \varphi(x)(f)$. Autrement dit, les éléments de A correspondent bijectivement aux morphismes de (X, A, φ) dans la droite affine (et non plus aux fonctions régulières ; un morphisme est une donnée plus riche que la fonction u sous-jacente).

Nos définitions montrent que l'étude des variétés algébriques affines est équivalente à celle des k-algèbres de type fini (c'est-à-dire engendrées par un nombre fini d'éléments). Les principaux résultats de cette théorie sont dus aux géomètres allemands de la première moitié du XXe siècle. Nous les citerons sans donner de démonstration complète.

Précisons d'abord que si B est un anneau, une B-algèbre A est dite de *type fini* si elle est engendrée en tant qu'algèbre par

un nombre fini d'éléments, tout élément de A s'exprimant comme fonction polynôme de ces générateurs. L'algèbre A est dite *finie* si elle est engendrée par un nombre fini d'éléments en tant que B-module : tout élément de A est combinaison linéaire des générateurs ; cela revient à dire que A est de type fini et entière sur B.

Lemme de normolisation d'Emmi Noether

Soit A une k-algèbre de type fini non nulle, engendrée par n éléments. Il existe un entier d et un homomorphisme *injectif* v : $k[T_1, T_2, \dots T_d] \rightarrow$ A, faisant de A une $k[T_1, T_2, \dots T_d]$-algèbre *finie*.

Géométriquement, v s'interprète comme un morphisme de la variété affine X qui correspond à A dans l'espace affine k^d ; les propriétés de v impliquent que ce morphisme est surjectif et « fini », c'est-à-dire qu'il fait de X une sorte de revêtement ramifié de k^d.

On peut démontrer ce lemme par récurrence sur le nombre n de générateurs de $A \simeq k[T_1, T_2, \dots, T_n]/\mathfrak{a}$; il est évident si $\mathfrak{a} = \{0\}$, en particulier si $n = 0$. Dans le cas contraire, on montre qu'il est possible de trouver un nouveau système de n générateurs dont le dernier est entier sur la sous-algèbre A' engendrée par les $n - 1$ premiers, de sorte que A est une A'-algèbre finie ; on applique alors l'hypothèse de récurrence à A'.

Par exemple, si $A = k[x, y]/(f)$ est l'algèbre de la courbe plane définie par l'équation $f(x, y) = 0$, on fait un changement de base dans k^2 de manière que l'axe des y ne soit pas une « direction asymptotique » de la courbe (c'est possible, car k est infini) ; l'équation de la courbe prend alors la forme :

$$y^r + a_1(x)y^{r-1} + \dots + a_r(x) = 0,$$

où les $a_i(s)$ sont des polynômes en x. Ainsi la classe de y dans A est entière sur $k[x]$, et la courbe apparaît comme un revêtement ramifié (de degré r) de l'axe des x.

Appliquons ce résultat en remplaçant A par le quotient $A/\mathfrak{m}$ où $\mathfrak{m}$ est un idéal maximal de A ; ce quotient est encore de type fini sur k, et c'est un corps. D'après le lemme, il contient une sous-algèbre B isomorphe à une algèbre de polynômes, sur laquelle il est une algèbre finie ; on en déduit aisément que B est elle-même un corps. et ensuite que $B \simeq k$. Comme k est algébriquement clos, son extension finie $A/\mathfrak{m}$ lui est isomorphe. Cela prouve que tout idéal maximal de A est le noyau d'un homomorphisme de A dans k, et permet d'établir une correspondance bijective entre les idéaux maximaux de A et les points de la variété algébrique affine associée à A. Pour développer la géométrie algébrique sur un corps non algébriquement clos, il est raisonnable de remplacer l'ensemble $\mathrm{Hom}_k(A, k)$ par l'ensemble des idéaux maximaux de A dans la définition des variétés algébriques affines ; cela revient à considérer, outre les « points rationnels sur k » de la variété, qui correspondent à des idéaux maximaux $\mathfrak{m}$ tels que $A/\mathfrak{m} \simeq k$, d'autres points correspondant a des idéaux maximaux m tels que $A/\mathfrak{m}$ soit une extension finie de k. *Nous* conservons un corps de base algébriquement clos pour rester plus près des notions intuitives.

Considérons une variété algébrique affine (X, A, φ) ; si $f \in A$ et $x \in X$, nous noterons $f(x)$ la valeur en x de la fonction sur X définie par f, c'est-à-dire $\varphi(x)(f)$. Pour qu'un élément de f de A soit inversible, il faut et il suffit que $f(x) \neq 0$ pour tout point x de X ; en effet, cette condition signifie que f n'appartient à aucun idéal maximal de A, et, d'après le théorème de

Krull, un élément non inversible appartient toujours à un idéal maximal.

Théorème des zéros de Hilbert

Soit (X, A, φ) une variété algébrique affine. Si $f \in$ A les conditions suivantes sont équivalentes :

(1) Pour tout point x de X, f(x) = 0.

(2) L'élément $1 - f$T est inversible dans l'algèbre de polynômes A[T].

(3) fest nilpotent, c'est-à-dire que l'une de ses puissances est nulle.

La condition (2) sert d'intermédiaire entre (1) et (3) ; elle s'interprète en disant que $1 - f$T ne s'annule pas sur la variété affine d'algèbre A[T] (qui n'est autre que X X k, cf. *infra*). Ainsi, on voit aisément que (1) implique (2). Pour voir que (2) implique (3), on note que $1 - f$T admet pour inverse $1 + f\mathrm{T} + f^2\mathrm{T}^2 + ...f^n\mathrm{T}^n +$ dans l'algèbre de séries formelles A [[T]], si cet inverse est un polynôme, fest nilpotent. Enfin, si f est supposé nilpotent, il en est de même de $f(x)$ pour tout point x ; or $f(x) \in k$, et dans un corps tout élément nilpotent est nul.

Autrement dit, l'ensemble n des éléments nilpotents de A est l'intersection des idéaux maximaux (c'est un idéal qu'on appelle le *nilradical* de A). Pour un anneau quelconque, le même raisonnement prouve que le nilradical est l'intersection des idéaux premiers ; ici, on voit, en appliquant le théorème précédent à $A/\mathfrak{p}$ où $\mathfrak{p}$ est un idéal premier de A, que tout idéal premier de A est une intersection d'idéaux maximaux (on dit que A est un anneau de Jacobson).

À tout idéal a de A, nous associerons l'ensemble V(a) C X des points x tels que f(x) = 0 pour tout $f \in$ a ; il suffit d'écrire cette condition pour un système de générateurs $(f_1, f_2, ... f_s)$ de a, et on peut dire que V(a) est un sous-ensemble algébrique de X. Avec l'algèbre $A/\mathfrak{a}$, il forme une variété algébrique affine à laquelle on peut appliquer le théorème précédent. On trouve ainsi que l'idéal I(V (a)) formé des élémentsfde A tels que $f(x) = 0$ pour tout x $\in$ V(a) est égal à la racine ta(a) de a c'est-à-dire à l'ensemble des éléments de A qui ont une puissance dans a ; en particulier, V(a) n'est pas vide si a $\neq$ A (c'est l'énoncé du *Nullstellensatz* de Hilbert).

3. Variétés algébriques

L'utilisation des fonctions régulières ne conduit à rien dans l'étude des ensembles algébriques projectifs, puisque l'anneau des fonctions régulières d'un tel ensemble est toujours réduit à k. On peut remplacer les fonctions régulières par les fonctions rationnelles. La théorie ainsi construite permet l'étude des propriétés conservées par une équivalence birationnelle ; elle a été développée à la fin du siècle dernier, principalement en Italie. Cette méthode interdit la distinction entre des ensembles algébriques birationnellement équivalents, même non isomorphes. Il faut donc chercher dans une autre direction pour obtenir une définition intrinsèque des variétés algébriques.

En localisant la notion de fonction régulière, on arrive à une notion adéquate. Commençons par munir les ensembles algébriques d'une topologie qui permette une telle localisation.

Topologie de Zariski

Si le corps de base est celui des nombres complexes, on peut essayer la topologie induite par celle de $\mathbf{C}^n$ ou P,,(C) (c'est la *topologie transcendante,* définie à partir du module d'un nombre complexe). Il est clair, en effet, que les applications réguliè-

res sont continues pour cette topologie ; par suite, les isomorphismes la conservent. Mais cela ne convient pas au cas d'un corps de base général.

On veut une topologie adaptée à l'étude des propriétés algébriques ; ainsi les propriétés algébriques (ou du moins beaucoup d'entre elles) doivent avoir une nature locale pour la topologie cherchée : par exemple, une fonction rationnelle définie en un point doit rester définie au voisinage de ce point. Or les ensembles exceptionnels où les propriétés algébriques considérées cessent d'être vraies sont définis par des équations polynomiales (l'annulation du dénominateur dans le cas de la définition d'une fonction rationnelle) ; ce sont eux-mêmes des ensembles algébriques. Ainsi, nous imposons la condition suivante : les ensembles algébriques doivent être fermés pour la topologie cherchée. Il se trouve qu'il existe sur k^n, ou sur $\mathbf{P}^n(k)$, une topologie bien déterminée pour laquelle les ensembles fermés sont les ensembles algébriques : on l'appelle *la topologie de Zariski.*

En effet, l'intersection d'une famille $(E_i)_{i \in I}$ d'ensembles algébriques est encore algébrique : si E_i est défini par les équations $f_{i_\lambda} = 0$, $\lambda \in \Lambda_i$, cette intersection est définie par l'annulation de tous les polynômes f_{i_λ} : c'est l'ensemble des zéros communs aux polynômes de l'idéal engendré par les f_{i_λ} et il suffit de prendre un ensemble fini de générateurs de cet idéal pour avoir un système d'équations de l'intersection (rappelons la propriété noethérienne de l'anneau des polynômes ; dans le cas projectif, il faut bien sûr prendre des générateurs homogènes). Considérons maintenant la réunion de deux ensembles algébriques, définis respectivement par les équations $f_1 = 0$, $f_2 = 0, \ldots, f_r = 0$ et $g_1 = 0$, $g_2 = 0$, ..., $g_s = 0$; c'est l'ensemble des zéros communs aux polynômes $f_i g_j$ ($1 \leqslant i \leqslant r$, $1 \leqslant j \leqslant s$), donc un ensemble algébrique. Enfin l'ensemble vide est algébrique, étant défini par l'équation 1 = 0. Ainsi les axiomes des fermés d'une topologie sont vérifiés par les ensembles algébriques (cf. TOPOLOGIE GÉNÉRALE). Dorénavant nous considérons k^n et $\mathbf{P}_n(k)$ comme munis de leurs topologies de Zariski, et tout ensemble algébrique affine ou projectif est muni de la topologie induite ; toute variété algébrique affine est de même munie de sa topologie de Zariski.

Si (X, A, φ) est une variété algébrique affine, une base d'ouverts de sa topologie de Zariski est formée des ensembles $D(f) = \{x \in X \mid f(x) \neq 0\}$, où $f \in A$. On obtient de même une base d'ouverts sur un ensemble algébrique projectif en considérant les ensembles où ne s'annule pas un polynôme homogène.

Il est clair que les applications régulières sont continues pour la topologie de Zariski : l'image réciproque d'un ensemble algébrique par une telle application est encore algébrique. Ainsi, deux ensembles algébriques isomorphes sont homéomorphes. La réciproque n'est pas vraie, puisque la représentation paramétrique $t \mapsto (t^2, t^3)$ de la cissoïde est un homéomorphisme de k sur cette courbe, qui n'est pourtant pas isomorphe à une droite (cf. chap. 1, *Applications régulières*) ; autrement dit, la topologie de Zariski est insuffisante pour caractériser les ensembles algébriques à isomorphisme près.

Anneaux locaux

À tout point x d'un ensemble algébrique X on peut associer l'anneau $\mathcal{O}_{X,x}$, des germes de fonctions rationnelles sur X définies **en** x ; un tel germe est une classe d'équivalence de fonctions rationnelles définies en x, pour la relation qui consiste à confondre

deux fonctions lorsqu'elles coïncident dans un voisinage de x. Ainsi un élément de $\mathcal{O}_{X,x}$ est représenté par une fraction P/Q dont le dénominateur Q ne s'annule pas en x ; dans le cas projectif, les polynômes P et Q sont homogènes de même degré. Pour qu'un tel élément soit inversible, il faut et il suffit que son numérateur P ne s'annule pas en x; autrement dit, l'ensemble des éléments non inversibles de $\mathcal{O}_{X,x}$ est l'idéal $\mathfrak{m}_x$, noyau de l'homomorphismef-f(x) à valeurs dans k, et c'est le seul idéal maximal de l'anneau. Un anneau qui possède un seul idéal maximal est dit local ; nous avons muni l'ensemble algébrique X d'un anneau local $\mathcal{O}_{X,x}$ pour chaque point x.

Si une application rationnelle $u : X \rightarrow Y$ est définie en un point x de l'ensemble algébrique X, il lui correspond un homomorphisme $g \mapsto g \circ u$ de l'anneau local $\mathcal{O}_{Y,u(x)}$ dans $\mathcal{O}_{X,x}$; cet homomorphisme applique l'idéal maximal du premier anneau dans celui du second, ce qu'on exprime en disant que c'est un *homomorphisme local.* Si (u, v) est une équivalence birationnelle entre X et Y, elle définit une bijection entre des ouverts de Zariski partout denses U et V de X et Y de manière que les anneaux locaux en des points correspondants de U et V soient isomorphes ; cela s'applique en particulier à un isomorphisme de X sur Y, avec U = X et V = Y.

On peut donner trois exemples :

- Les fonctions rationnelles sur la droite k forment le corps $k(x)$ (corps des fractions de l'anneau des polynômes $k[x]$). Comme les voisinages ouverts au sens de Zariski d'un point a sont des complémentaires d'ensembles finis, deux fonctions rationnelles qui coïncident au voisinage de a sont identiques, et l'anneau local de a n'est autre que l'ensemble des fractions dont le dénominateur n'est pas nul en a.

- Il est facile de voir que les fonctions rationnelles sur la courbe d'équation $y^2 = x^3$ s'écrivent d'une manière unique sous la forme R(x) + y S(x), où R et S sont des éléments de $k(x)$; ainsi l'ensemble des fonctions rationnelles sur la cissoïde est un corps, extension quadratique de $k(x)$ engendrée par un élément y qui vérifie $y^2 = x^3$. L'anneau local d'un point (a, b) s'identifie à l'ensemble des éléments R(x) + y S(x) tels que R et S soient définies en a. Comme les applications $u : t \mapsto (t^2, t^3)$ (application régulière) et $v = y/x$ (application rationnelle définie sauf en (0, 0)) forment une équivalence birationnelle entre la droite et la cissoïde X, les anneaux locaux sur X aux points $(a, b) \neq (0, 0)$ sont isomorphes à ceux des points correspondants sur la droite. Au contraire, l'anneau local, $\mathcal{O}_{X,(0,0)}$ n'est pas intégralement clos (il ne contient pas l'élément entier y/x), tandis que tous les anneaux locaux sur la droite sont intégralement clos. On voit donc que le point de rebroussement de la cissoïde se manifeste dans les propriétés des anneaux locaux (fig. 2).

Un raisonnement analogue permet d'étudier le cône d'équation $z^2 = x^2 + y^2$ dans k^3. Les fonctions rationnelles sur cette surface forment un corps, extension quadratique de $k(x, y)$; elles s'écrivent $R(x, y) + zS(x, y)$. L'anneau local de l'origine est formé de ceux de ces éléments pour lesquels les fractions R et S sont définies en (0, 0), et son idéal maximal est défini par la condition R(O,O) = 0.

Faisceau structural

La connaissance de la topologie et des anneaux locaux sur un ensemble algébrique est insuffisante pour caractériser cet ensemble à isomorphisme près ; en particulier, elle ne permet pas de reconstituer l'algèbre

des fonctions régulières sur l'ensemble. Nous allons remplacer les anneaux locaux par une structure plus riche.

Considérons un ouvert de Zariski U dans un ensemble algébrique X. Nous dirons qu'une application de U dans k est une fonction régulière dans U si son germe en chaque point x de U appartient à l'anneau local $\mathcal{O}_{X,x}$; l'ensemble des fonctions régulières dans U forme un anneau (et même une k-algèbre), que nous noterons O,(U). Notre terminologie n'est pas contradictoire, car $\mathcal{O}_X(X)$ est précisément l'ensemble des fonctions régulières sur X, en vertu du théorème :

Soient X un ensemble algébrique et f : X $\rightarrow k$ une application de X dans k. On suppose que pour tout point x de X il existe un voisinage V_x de x et une fonction rationnelle r_x définie dans V_x et telle que $f|_{V_x} = r_x$. Alors f est une fonction régulière.

Ce résultat se démontre ainsi, dans le cas où X est affine : on peut supposer que les voisinages V_x sont de la forme $D(g_x)$ (ensemble des points où g_x ne s'annule pas), avec des fonctions régulières g,. Comme les $D(g_x)$ recouvrent X, l'idéal engendré par les fonctions g_x est l'anneau A de toutes les fonctions régulières sur X (sinon on pourrait trouver un idéal maximal contenant les g_x, donc un point de X où toutes ces fonctions s'annulent) ; par suite, 1 est combinaison linéaire d'un nombre fini de fonctions g_x soient g_1, g_2, ... g_s et X = $D(g_1) \cup D(g_2) \cup \cup D(g_s)$. Dans l'ouvert $D(g_i)$, f coïncide avec une fonction rationnelle r_i définie dans $D(g_i)$; ici nous avons besoin d'un lemme : toute fonction rationnelle définie dans D(g,) peut s'écrire sous la forme h/g_i^m où h est une fonction régulière et m un entier (car g_i s'annule sur l'ensemble des zéros du dénominateur, donc, par le *Nullstellensatz*, ce dénominateur divise une puissance de g_i). Nous pouvons donc trouver un entier m et des fonctions régulières $h_1, h_2, \dots, h_s$ tels que $g_i^m f$ coïncide avec h_i dans $D(g_i)$; dans $D(g_i) \cap D(g_j)$ = D(g,g,), les fonctions $g_i^m h_j$ et $g_j^m h_i$ sont égales, et par suite leur différence est annulée par $g_i g_j$. Il existe des fonctions régulières $f_1, f_2, \dots, f_s$ telles que $f_1 g_1^{m+1} + f_2 g_2^{m+1} + \dots + f_s g_s^{m+1} = 1$, car les $D(g_i^{m+1})$ recouvrent X; on peut alors vérifier que f coïncide avec $f_1 g_1 h_1 + f_2 g_2 h_2 + \dots + f_s g_s h_s$ dans chaque D(g,), donc lui est égale.

On appelle *faisceau d'anneaux* sur un espace topologique X un couple (A, p) du type suivant : A associe à chaque ouvert U de X un anneau A(U) et ρ associe à chaque couple (U, V) d'ouverts tels que U C V un hornorphisme ρ_{UV} : A(V) $\rightarrow$ A(U) appelé *restriction* de V à U. On impose à ces données les conditions suivantes :

(1) Pour tout ouvert U, l'homomorphisme ρ_{UU} est l'application identique de A(U). Si UCVCW, on a $\rho_{UW} = \rho_{UV} \circ \rho_{VW}$.

(2) Pour tout ouvert U et tout recouvrement $(U_i)_{i \in I}$ de U par des ouverts, l'application $f \mapsto (\rho_{U_iU}(f))$ est une bijection de A(U) sur l'ensemble des familles :

$$(f_i) \in \prod_{i \in I} A(U_i),$$

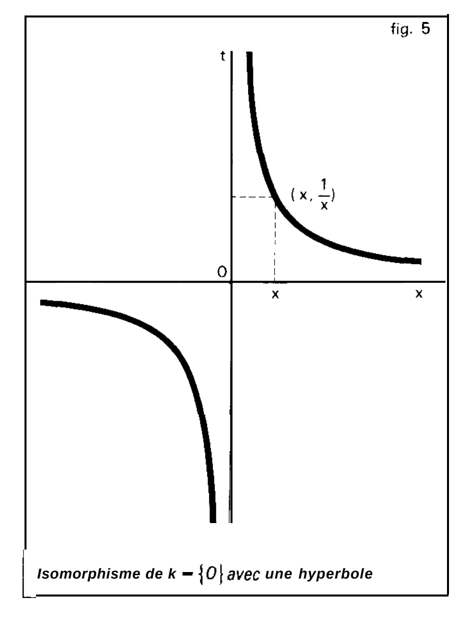
telle que les restrictions de f_i et f_j à $U_i \cap U_j$ soient égales pour tout couple (i, j) d'indices. Cela signifie qu'un élément de A(U) est connu quand on connaît ses restrictions à tous les U_i, et donne la condition pour que des éléments des A(U,) proviennent par restriction d'un même élément de A(U).

La *fibre* A, du faisceau en un point Y de X est la limite inductive des A(U) lorsque U parcourt l'ensemble filtrant des voisinages ouverts de x ; pour construire cette

limite inductive, on identifie des éléments de A(U) et A(V), où U et V sont des voisinages ouverts de x, si leurs restrictions à un voisinage W C U $\cap$ V sont égales ; les classes d'équivalence ainsi définies sont les éléments de A,.

11 est clair que les anneaux de fonctions 0x(U) que nous avons définis sur un ensemble algébrique X forment un faisceau d'anneaux (avec la notion habituelle de restriction). Les fibres sont les anneaux locaux $\mathcal{O}_{X,x}$. Considérons une application régulière u : X-t Y d'ensembles algébriques ; elle est continue pour les topologies de Zariski, de sorte que si V est un ouvert de Y, u^{-1}(V) est un ouvert de X. On a alors un homomorphisme g $\mapsto$ g $\circ$ u de $\mathcal{O}_Y$(V) dans $\mathcal{O}_X(u^{-1}(V))$. D'une manière générale, on appelle *espace localement annelé* un espace topologique X muni d'un faisceau d'anneaux (A, p) dont les fibres sont des anneaux locaux. Un morphisme d'espaces localement annelés de (X, (A, p)) dans (Y, (B, σ)) est un couple (u, v) où u est une application continue de X dans Y et v associe à chaque ouvert V de Y un homomorphisme v_V : B(V)$\longrightarrow$A(u^{-1}(V)) de manière compatible avec les restrictions ρ et σ. On impose de plus que, pour tout point x de X, l'homomorphisme $v_x : B_{u(x)} \longrightarrow$ A, déduit de v soit un homomorphisme local. Il est facile de définir le morphisme composé de deux morphismes, puis la notion d'isomorphisme entre espaces localement annelés.

Soit U un ouvert d'un espace localement annelé (X, (A, p)). La topologie induite et le faisceau induit A$|_U$: U' $\mapsto$ A(U'), pour U' ouvert contenu dans U, font de U un espace localement annelé ; on dit que c'est un sous-espace ouvert de X. Par exemple, si X est un ensemble algébrique affine et si f est une fonction régulière sur X, l'ensemble D(f) des points de X oùf'n'est pas nulle, muni de la structure d'espace annelé induite par celle de X, est un sous-espace ouvert de X. Si X est défini, dans k^n, par les équations $g_1 = 0$, g_2 = 0, ... g_s = 0, et si f est la restriction à X d'un polynôme $f \in k[x_1, x_2, \ldots x_n]$, alors D($f$) est la projection sur k^n de l'ensemble algébrique Y C k^{n+1} = k^n X k défini par les équations g, = 0, $g_2 = 0$, ..., $g_s = 0$ et $1 - tf = 0$. La projection : (x, t) $\mapsto$ x définit un isomorphisme d'espaces localement annelés de Y sur D(f). Par exemple, le complémentaire de 0 dans la droite k est isomorphe, comme espace localement annelé, à l'hyperbole d'équation xt = 1 dans le plan des (x, t) (fig. 5).



Isomorphisme de k – {O} avec une hyperbole

Nous pouvons aussi munir D(f) d'une structure de variété algébrique affine, en transportant celle de Y. Soit A = $k[x_1, x_2, x_n]/\mathfrak{a}$ l'algèbre de la variété affine X, où a est l'idéal engendré par g_1, g_2, ..., g_s. L'algèbre de Y est

$k[x_1, x_2, ..., x_n, t]/\mathfrak{b}$ où b est engendré par $g_1, g_2, ..., g_s$ et $1 - tf$; elle est isomorphe à $A_f = A[t]/(1 - tf')$, en désignant par f' la classe de f dans A. Dorénavant, si (X, A, φ) est une variété algébrique affine et si g est un élément de A, nous munirons l'ouvert $D(f)$ des points de X où f n'est pas nul d'une structure de variété algébrique affine au moyen de l'algèbre $A_f = A[t]/(1 - tf)$; les éléments de cette algèbre peuvent s'écrire comme des fractions du type h/f^r où $h \in A$ (notons qu'une telle fraction est nulle dès que h est annulé par une puissance de f, soit $f^m h = 0$; cela n'exige pas que h soit nul). On peut montrer qu'il existe sur X un faisceau $\mathcal{O}_X$ et un seul qui prennent la valeur A_f dans l'ouvert $D(f)$ pour tout $f \in A$. Si U est un ouvert de X, les éléments de $\mathcal{O}_X(U)$ ne sont pas des fonctions dans U, mais chacun d'eux définit une fonction (régulière) dans U : cette fonction est nulle si. et seulement si, l'élément considéré est nilpotent. La fibre de $\mathcal{O}_{X,x}$ en un point x est un anneau local, dont les éléments s'écrivent comme des fractions h/g avec $h, g \in A$ et $g(x) \neq 0$; ainsi (X, $\mathcal{O}_X$) est un espace localement annelé.

Si (u, v) : (X, A, φ) ⟶ (Y, B, ψ) est un morphisme de variétés algébriques affines, on en déduit aisément un morphisme $(u, \tilde{v})$: (X, $\mathcal{O}_X$) ⟶ (Y, $\mathcal{O}_Y$) en définissant $\tilde{v}_{D(g)}$, pour $g \in B$, comme l'homomorphisme de B_g dans $A_{v(g)}$ qui transforme h/g^r en $v(h)/v(g)^r$. Inversement, si (u, w) : (X, $\mathcal{O}_X$) ⟶ (Y, $\mathcal{O}_Y$) est un morphisme d'espaces localement annelés, (u, w_Y) est un morphisme de variétés affines ; on a $(\tilde{v})_Y = v$ et $(\tilde{w}_Y) = w$. Cela implique en particulier que deux variétés algébriques affines sont isomorphes si, et seulement si, elles sont isomorphes comme espaces localement annelés. Dans la suite, nous considérerons les variétés algébriques affines comme des espaces localement annelés.

Considérons maintenant un ensemble algébrique X défini dans l'espace projectif $P_n(k)$ par des équations homogènes $g_1 = 0$, $g_2 = 0$, ..., $g_s = 0$, et un polynôme f homogène de degré d par rapport aux coordonnées homogènes $x_0, x_1, ..., x_n$. Soit $D_+(f)$ l'ensemble ouvert des points de X où f n'est pas nul, muni de la structure annelée induite par celle de X ; il est isomorphe à un ensemble algébrique affine. On le voit tout de suite si $f = x_0$, en utilisant l'isomorphisme connu de k^n sur le complémentaire de l'hyperplan d'équation $x_0 = 0$ dans $P_n(k)$; si $d = 1$, un changement de coordonnées nous ramène à ce cas facile. Le cas général se ramène au cas où $d = 1$ en identifiant $P_n(k)$ à la variété de Veronese $V_{n,d}$ (cf. chap. 1).

Cela permet encore de munir $D_+(f)$ d'une structure de variété algébrique affine, dont l'algèbre est obtenue ainsi : soit $\mathfrak{a}$ l'idéal homogène de polynômes engendré par $g_1, g_2, ..., g_s$, et soit $A = k[x_0, x_1, ..., x_n]/\mathfrak{a}$ l'algèbre graduée quotient ; désignons par $A^{(d)}$ la sous-algèbre de A formée des éléments dont le degré est un multiple de d; l'algèbre de $D_+(f)$ est $A_{(f)} = A^{(d)}/(f - 1)$ et ses éléments s'écrivent comme des fractions h/f^r où h est un élément de degré rd de A. Il existe alors un faisceau d'anneaux bien déterminé sur X qui prend la valeur $A_{(f)}$ sur $D_+(f)$; sa fibre en un point x est un anneau local, dont les éléments sont les fractions du type h/g avec $h, g \in A$ homogènes de même degré et g non nul en x. Nous désignerons ce faisceau par $\mathcal{O}_X$, et nous dirons que l'espace localement annelé (X, $\mathcal{O}_X$) est une variété algébrique projective. On peut montrer que pour tout morphisme d'espaces localement annelés (u, w) : (X, $\mathcal{O}_X$) ⟶ (Y, $\mathcal{O}_Y$) entre des variétés algébriques projectives,

l'application u est régulière (on commence par observer que la restriction de u à tout ouvert affine de la forme $D_+(f) \subset X$ est régulière).

La notion d'espace localement annelé nous a permis de traiter de manière analogue les variétés algébriques affines ou projectives. De plus, toute variété algébrique projective X C P,,(k) peut être recouverte par des ouverts U, qui sont des variétés affines pour la structure induite ; on peut même prendre des U, en nombre fini, par exemple $U_i = D_+(x_i)$ (x_i coordonnées homogènes). D'une manière générale, nous dirons qu'un espace localement annelé $(X, \mathcal{O}_X)$ est une *variété algébrique* s'il existe un recouvrement (U,) de X par des ouverts qui sont isomorphes (pour la structure localement annelée induite) à des variétés algébriques affines.

4. Propriétés élémentaires

Tout ouvert U d'une variété algébrique X, muni de la structure annelée induite, est une variété algébrique ; on dit que c'est une *sous-variété ouverte* de X.

Considérons un faisceau d'idéaux $\mathfrak{J}$ de $\mathcal{O}_X$ (c'est-à-dire un faisceau tel que J(U) soit un idéal de 0,(U) pour tout ouvert U, les opérations de restriction de $\mathfrak{J}$ étant induites par celles de $\mathcal{O}_X$) ; on peut définir un faisceau quotient $\mathcal{O}_X/\mathfrak{J}$, dont la fibre en un point x quelconque de X est $\mathcal{O}_{X,x}/\mathfrak{J}_x$. Si $\mathfrak{J}$ est « localement de type fini », le support Y de ce faisceau quotient, c'est-à-dire l'ensemble des points x où sa fibre n'est pas nulle, est une partie fermée de X. On peut considérer $\mathcal{O}_X/\mathfrak{J}$ comme un faisceau sur Y, et Y muni de ce faisceau est une variété algébrique ; on dit que c'est une *sous-variété fermée* de X. Par exemple, les variétés algébriques affines (resp. projectives) peuvent être considérées comme des sous-variétés fermées d'un espace k^n (resp. $\mathbf{P}_n(k)$).

Si Y est un fermé quelconque de X, le faisceau d'idéaux $\mathfrak{J}_Y$ formé des f qui s'annulent sur Y est localement de type fini et $\mathcal{O}_X/\mathfrak{J}_Y$ a pour support Y, d'où sur Y une structure de sous-variété fermée de X. Ce n'est pas la seule possible, mais on peut la caractériser par le fait qu'elle est *réduite*, c'est-à-dire que son faisceau structural ne comporte pas d'éléments nilpotents non nuls (il s'interprète comme un faisceau de fonctions sur Y). En particulier, si Y = X, on définit une sous-variété fermée X_{red} de X qui est réduite et a même espace topologique sous-jacent que X.

Sur l'ensemble produit X X Y de deux variétés algébriques, on peut définir une structure de variété algébrique munie de morphismes p : X X Y - X et $q : X \times Y \rightarrow Y$, de manière que les morphismes u d'une variété Z dans X X Y correspondent bijectivement aux couples $(p \circ u, q \circ u)$ de morphismes de Z dans X et Y respectivement. En général la topologie de X X Y est strictement plus fine que la topologie produit. Par exemple k"' X k" $\simeq k^{m+n}$; le produit de deux variétés affines est par suite une variété affine. De même, le produit de deux variétés projectives est projective ; en effet, $\mathbf{P}_m(k)$ X P,,(k) s'identifie à une sous-variété fermée de P,.(k), avec $\mathrm{Y} = mn + m + n$, au moyen du *morphisme de Segre* qui transforme un couple $(x, y) \in \mathbf{P}_m(k) \times \mathbf{P}_n(k)$ en le point de coordonnées homogènes $x_i y_j$ $(0 \leqslant i \leqslant m, 0 \leqslant j \leqslant n)$ où les x_i sont les coordonnées homogènes de x et les y_j celles de y. Par exemple, pour $m = n = 1$, le morphisme de Segre P,(k) X P,(k) $\rightarrow$ P,(k) identifie le produit de deux droites projectives à la quadrique d'équation homogène $zt = xy$

(cf. chap. 1, exemple des *paraboloïdes* et fig. 3) ; lorsque a décrit P,(k), l'image de {a} X P,(k) décrit l'un des systèmes de génératrices rectilignes de la quadrique, tandis que l'image de P,(k) X $\{a\}$ décrit l'autre système.

On dit qu'une variété algébrique X est *séparée* si la diagonale $\Delta_X = \{(x, x)$ $x \in \Gamma X\}$ est une partie fermée de X X X (comme la topologie de X X X est plus fine que la topologie produit, cela ne signifie pas en général que la topologie de X est séparée). Une sous-variété (ouverte ou fermée) d'une variété séparée est aussi séparée ; de même, un produit de variétés séparées est séparé. L'espace k^n et l'espace P,(k) sont séparés, donc les variétés algébriques affines ou projectives sont séparées.

La propriété noethérienne des algèbres de type fini se traduit dans le fait que toute suite décroissante de fermés d'une variété algébrique affine est stationnaire. Il en résulte que tout sous-espace d'une telle variété est quasi compact, c'est-a-dire vérifie l'axiome de Borel-Lebesgue (cf. TOPOLOGIE GÉNÉRALE). Les variétés algébriques qui sont quasi compactes sont celles qui admettent un recouvrement fini par des ouverts affines ; il en est ainsi pour les variétés affines ou projectives. Les variétés algébriques au sens de Serre sont supposées réduites, séparées et quasi compactes.

Une variété non vide qui n'est pas réunion de deux fermés strictement plus petits est dite *irréductible* ; il revient au même de dire que tout ouvert non vide est partout dense. L'espace k^n et l'espace P,(k) sont irréductibles. Toute variété algébrique quasi compacte est réunion d'un nombre fini de *composantes irréductibles,* qui sont ses parties fermées irréductibles maximales.

La *dimension* d'une variété algébrique X est définie comme la borne supérieure de l'ensemble des entiers n pour lesquels il existe une suite strictement croissante de (F,, F,, ... F,) de fermés irréductibles de X ; c'est un nombre fini si X est quasi compacte. Si X est une variété intègre, c'est-à-dire irréductible et réduite, on peut définir des *fonctions rationnelles* sur X ; ces fonctions forment un corps K(X), extension de type fini de *k* dont le degré de transcendance est égal à la dimension de X (cf. CORPS). Ainsi l'espace k^n et l'espace P,(k) sont de dimension n, leur corps de fonctions rationnelles étant $k(x_1, x_2, \ldots . x,)$.

Si *k* est le corps des nombres complexes, on peut considérer $k^n = \mathbf{C}^n$, comme une variété analytique complexe. Toute variété algébrique affine X, sous-variété fermée de $\mathbf{C}^n$ peut alors être munie d'une structure d'espace analytique X^{an} sous-espace analytique fermé de $\mathbf{C}^n$ (la topologie de X^{an} est induite par la topologie transcendante de $\mathbf{C}^n$ et son faisceau structural est quotient du faisceau des fonctions analytiques). Considérons maintenant une variété algébrique X sur C, recouverte par des ouverts affines U, ; les structures analytiques U_i^{an} se recollent et définissent sur X une structure d'espace analytique X^{an}, de même dimension que X. Un grand nombre de propriétés de géométrie algébrique concernant X se traduisent par des propriétés analytiques de X^{an}.

Par exemple, on peut définir la notion de *point régulier* d'une variété algébrique X ; si le corps de base est C, pour qu'un point x de X soit régulier, il faut et il suffit que X^{an} soit une variété analytique sans singularité au voisinage de x. Ainsi, tout point de $\mathbf{C}^n$ ou de P,,(C) est régulier, mais 0 n'est pas un point régulier de la cissoïde représentée dans la figure 2 (on dit que c'est un point singulier). L'ensemble des

points réguliers d'une variété algébrique est ouvert, et il est partout dense si la variété est réduite. L'anneau local en un point régulier est factoriel (cf. ANNEAUX COMMUTATIFS).

Comme autre exemple, considérons la notion de *variété algébrique complète* (définie sur un corps de base arbitraire). Dans le cas où le corps de base est C, pour que X soit complète il faut et il suffit que X^{an} soit *compact.* On voit ainsi que P,,(C) et que les variétés algébriques projectives sont des variétés complètes. Au contaire, les variétés affines de dimension non nulle ne sont pas complètes.

Dans la suite de cet article, nous remplaçons les définitions de certaines notions de géométrie algébrique par leurs traductions transcendantes, en nous restreignant au cas où le corps de base est C ; c'est le point de vue de la géométrie italienne du siècle dernier (on démontrait alors les théorèmes par des méthodes transcendantes). Nous pourrons ainsi énoncer un certain nombre de résultats sans être entraînés à des développements trop longs ; cependant, ces résultats conservent leur sens et leur validité avec un corps de base général.

5. Morphismes finis. Normalisation et désingularisation

On dit qu'un morphisme $f = (u, v) : X \to Y$ de variétés algébriques affines est *fini* si v_Y : B $\to$ A fait de A une B-algèbre finie (A désigne l'algèbre de X et B celle de Y). Plus généralement, un morphisme f : X $\to$ Y entre des variétés algébriques quelconques est dit *fini* s'il existe un recouvrement de Y par des ouverts affines U_i tels que les ouverts $f^{-1}(U_i)$ de X soient affines et que les restrictions :

$$f_{|f^{-1}(U_i)} : f^{-1}(U_i) \to U_i$$

soient finies. Un morphisme fini transforme les fermés de X en fermés de Y, et, pour tout point y de Y, la fibre $f^{-1}(y)$ est finie et discrète. Le lemme de normalisation de Noether, énoncé au chapitre 2, signifie que si X est une variété algébrique affine, il existe un morphisme fini et surjectif de X sur un espace k^d ; l'entier d est égal à la dimension de X. Si X est une sous-variété fermée d'une variété Y. le morphisme d'injection X $\to$ Y est fini.

Soit f : X $\to$ Y un morphisme de variétés algébriques tel que pour tout point y de Y la fibre $f^{-1}(y)$ soit finie et discrète. En général f n'est pas fini, mais, si on le suppose séparé (cela signifie que $\Delta_X = \{(x, x) \quad x \in X\}$ est fermé dans le « produit fibré » $X \times_Y X$; c'est toujours vrai si X est séparé), on peut montrer qu'il existe une variété algébrique X' dans laquelle X se plonge comme sous-variété ouverte de manière que f se prolonge en un morphisme fini de X' dans Y. Ce résultat, profond et difficile, est connu sous le nom de *théorème principal* de Zariski.

Un point x d'une variété algébrique X est dit *normal* si l'anneau local $\mathcal{O}_{X,x}$ est intègre et intégralement clos. Sur le corps des complexes, cela revient à dire que x est un point normal de X^{an}, c'est-à-dire que toute fonction analytique définie seulement aux points réguliers voisins de x et bornée se prolonge en une fonction analytique définie dans un voisinage de x. Un point régulier est normal ; sur une courbe la réciproque est vraie, mais pas en dimension plus grande. Par exemple, le sommet 0 du cône d'équation $z^2 = x^2 + y^2$ dans k^3 est normal sans être régulier (cf. chap. 3 et

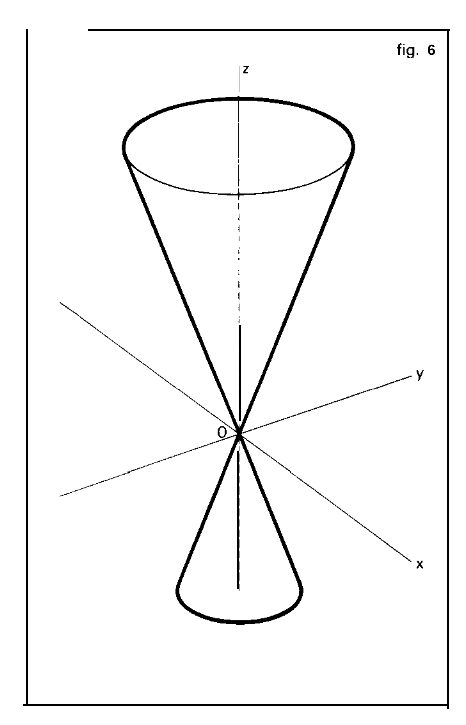
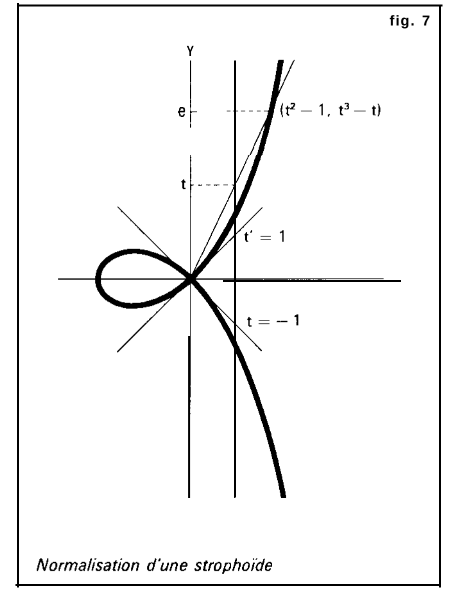


fig. 6). En un point normal x, chaque composante irréductible de l'ensemble S des points singuliers est de codimension au moins 2 ; par exemple, si X est une courbe, x a un voisinage qui ne contient aucun point singulier ; si X est une surface, il existe un voisinage de x qui ne contient pas d'autre singularité que x.

Soit X une variété algébrique affine intègre, d'algèbre A. On démontre que la clôture intégrale $\tilde{A}$ de A (ensemble des éléments du corps K(X) qui sont entiers sur A ; cf. ANNEAUX COMMUTATIFS) est finie sur A ; il lui correspond une variété algébrique affine $\tilde{X}$ dont tous les points sont normaux, munie d'un morphisme fini $\tilde{X} \rightarrow X$. On généralise ce résultat en associant à toute variété réduite X une variété normale $\tilde{X}$ et un morphisme $\pi : \tilde{X} \rightarrow X$ fini et surjectif qui induit un isomorphisme de $\tilde{X} - \pi^{-1}(Z)$ sur $X - Z$ en désignant par Z l'ensemble des points de X qui ne sont pas normaux (c'est un fermé sans point intérieur) ; les variétés $\tilde{X}$ et X sont birationnellement équivalentes. Lorsque X est une courbe, $\tilde{X}$ est une courbe sans point singulier ; par exemple si X est unicursale et affine, $\tilde{X}$ est isomorphe à k, et π donne une représentation paramétrique de X (cf. fig. *2* et fig. 7).



Normalisation d'une strophoïde

11 est beaucoup plus difficile d'éliminer d'une manière analogue les singularités des variétés de dimension plus grande. Ce problème a été résolu par S. Abhyankar pour les surfaces et par H. Hironaka pour les variétés de dimension quelconque sur un corps de caractéristique 0 ; on associe à une variété algébrique X une variété X' sans point singulier et un morphisme surjectif $\varphi : X' \rightarrow X$ (non fini en général) qui est une équivalence birationnelle et un isomorphisme de $X' - \varphi^{-1}(S)$ sur $X - S$ (S désignant l'ensemble des points singuliers de X).

6. Faisceaux cohérents et cohomologie

Les méthodes cohomologiques sont, comme dans la théorie des espaces analytiques, un des outils les plus puissants de la géométrie algébrique. La topologie de Zariski permet de développer une théorie de la cohomologie à valeur dans les faisceaux algébriques cohérents sur les variétés algébriques.

Considérons une variété algébrique $(X, \mathcal{O}_X)$. On définit un faisceau de $\mathcal{O}_X$-modules $\mathcal{F}$, *ou faisceau algébrique* sur X, en associant à chaque ouvert U de X un $\mathcal{O}_X(U)$-module $\mathcal{F}(U)$ et en se donnant, pour $U \subset V$, des opérations de restriction $\mathcal{F}(V) \rightarrow \mathcal{F}(U)$ « semi-linéaires » relativement à celles de $\mathcal{O}_X$. Ces données sont soumises à des axiomes (1) et (2) analogues à ceux énoncés au chapitre 3 pour les faisceaux d'anneaux. La fibre :

$$\mathcal{F}_x = \lim_{U \ni x} \mathcal{F}(U),$$

en un point x de X est un $\mathcal{O}_{X,x}$-module. Un morphisme u d'un faisceau algébrique $\mathcal{F}$ dans un autre $\mathcal{G}$ est la donnée, pour chaque ouvert U de X, d'une application $\mathcal{O}_X(U)$-linéaire de u_U de $\mathcal{F}(U)$ dans $\mathcal{G}(U)$; si $U \subset V$, on impose que u_U et u_V soient compatibles avec les restrictions de V à U dans $\mathcal{F}$ et $\mathcal{G}$. Par exemple, si X est une variété algébrique affine, d'algèbre A, et si M est un A-module, on peut lui associer un faisceau algébrique $\tilde{M}$ dont la valeur dans un ouvert $D(f)$ de X avec $f \in A$, est $M_f = M[t]/(1 - tf)M[t]$ (les éléments de M_f se représentent comme des fractions m/f^r avec $m \in M$) ; à toute application A-linéaire $v : M \rightarrow N$ entre A-modules correspond un morphisme $\tilde{v} : \tilde{M} \rightarrow \tilde{N}$ de faisceaux algébriques ; le faisceau $\mathcal{O}_X$ est égal à $\tilde{A}$.

On peut montrer que le fait, pour un faisceau algébrique $\mathcal{F}$ sur la variété affine X, d'être isomorphe à un faisceau $\tilde{M}$, où M est un A-module, est une propriété locale. On peut donc définir une propriété correspondante sur toute variété algébrique. Nous dirons qu'un faisceau algébrique $\mathcal{F}$ sur une variété algébrique X est *cohérent* s'il existe un recouvrement de X par des ouverts affines U_i d'algèbres A_i et des A_i-modules de type fini M_i tels que la restriction de $\mathcal{F}$ à U_i soit isomorphe à $\tilde{M}_i$ pour tout i. Par exemple, si $\pi : E \rightarrow X$ est un fibré vectoriel algébrique localement trivial, on lui associe un faisceau cohérent $\mathcal{E}$ dont la valeur dans un ouvert U est l'ensemble des morphismes $\sigma : U \rightarrow E$ tels que $\pi \circ \sigma$ soit l'injection canonique de U dans X (« sections de E au-dessus de U ») ; le faisceau $\mathcal{E}$ est même localement libre, c'est-à-dire qu'on peut choisir les ouverts U_i de manière que $\mathcal{E}|_{U_i} \simeq \tilde{M}_i$ où M_i est un module libre de type fini.

À un faisceau algébrique cohérent $\mathcal{F}$ sur une variété algébrique X, on associe les groupes de cohomologie $H^n(X, \mathcal{F})$ ($n \in \mathbb{N}$) ; on a $H^0(X, \mathcal{F}) = \mathcal{F}(X)$. Si X est affine, $H^n(X, \mathcal{F}) = 0$ pour $n \geqslant 1$ et pour tout faisceau cohérent $\mathcal{F}$; cette propriété (analogue au théorème B de Cartan en géométrie analytique) caractérise les variétés algébriques affines.

Pour étudier la cohomologie d'une sous-variété fermée de l'espace projectif $\mathbf{P}_r(k)$, on utilise le faisceau fondamental $\mathcal{O}(1)$ ainsi défini. Désignons par π l'application canonique $k^{r+1} - \{0\}$ sur $\mathbf{P}_r(k)$, et considérons la sous-variété algébrique E de $\mathbf{P}_r(k) \times k^{r+1}$ définie par les conditions $\xi = 0$ ou $\xi \neq 0$ et $\pi(\xi) = x$ ($x \in \mathbf{P}_r(k)$, $\xi \in k^{r+1}$), et obtenue en faisant « éclater 0 » dans k^{r+1} (cf. fig. 3 pour le cas $r = 1$) ; c'est un sous-fibré vectoriel de rang 1 du fibré trivial $\mathbf{P}_r(k) \times k^{r+1}$ de base $\mathbf{P}_r(k)$ et

de fibre k^{r+1}. Le faisceau $\mathcal{O}(1)$ des sections du fibré dual E* est localement libre de rang 1 et $H^0(\mathbf{P}_r, \mathcal{O}(1))$ s'identifie à l'ensemble des formes linéaires sur k^{r+1} ; sa puissance tensorielle n-ième, notée $\mathcal{O}(n)$, est encore un faisceau localement libre de rang 1 et $H^0(\mathbf{P}_r, \mathcal{O}(n))$ s'identifie à l'ensemble des polynômes homogènes de degré n sur k^{r+1}. Si X est une sous-variété fermée de $P_r(k)$, nous désignerons par $\mathcal{O}_X(1)$ le faisceau des sections du dual de $E_X = X \times_{\mathbf{P}_r} E$, et par $\mathcal{O}_X(n)$ la puissance tensorielle n-ième de ce faisceau ; ce sont des faisceaux localement libres de rang 1 sur X ; enfin, si $\mathcal{F}$ est un faisceau algébrique cohérent sur X, le produit tensoriel :

$$\mathcal{F} \otimes_{\mathcal{O}_X} \mathcal{O}_X(n)$$

est noté $\mathcal{F}(n)$. J.-P. Serre a démontré que, pour un faisceau cohérent $\mathcal{F}$ donné, il existe un entier n_0 tel que $H^q(X, \mathcal{F}(n)) = 0$ pour $n \geqslant n_0$ et $q \geqslant 1$. De plus $H^q(X, \mathcal{F})$ est un k-espace vectoriel de dimension finie pour tout q et s'annule pour $q >$ dim X. Plus généralement $H^q(X, \mathcal{F})$ est de dimension finie sur k pour tout q lorsqu'on suppose seulement que X est complète.

Lorsque k = C, à tout faisceau algébrique cohérent $\mathcal{F}$ sur une variété algébrique X correspond un faisceau analytique cohérent $\mathcal{F}^{an}$ sur l'espace analytique X^{an}. Si X est projective, on a $H^n(X, \mathcal{F}) = H^n(X^{an}, \mathcal{F}^{an})$ pour tout n ; on peut en déduire que tout sous-espace analytique fermé Y de $\mathbf{P}_r(\mathbf{C})$ est de la forme X^{an} où X est une variété algébrique *(théorème de Chow)* et que tout faisceau analytique cohérent $\mathcal{G}$ sur Y est de la forme $\mathcal{F}^{an}$ où $\mathcal{F}$ est un faisceau algébrique cohérent bien déterminé sur X.

Considérons une variété algébrique X sur le corps des nombres complexes. Les propriétés topologiques de X^{an} (topologie transcendante) se manifestent dans les groupes de cohomologie à valeur dans des faisceaux constants, par exemple dans les groupes $H^n(X^{an}, Z)$. Les groupes correspondants calculés sur X avec la topologie de Zariski n'ont pas les propriétés raisonnables attendues, car la topologie de Zariski est trop grossière ; par exemple, il existe des « revêtements » Y → X (ou Y^{an} sur un revêtement de X^{an}) qui ne sont pas localement triviaux pour la topologie de Zariski. Cependant Grothendieck est parvenu à construire une théorie cohomologique valable sur un corps de base général et donnant de bons groupes de cohomologie à valeur dans des faisceaux constants *(cohomologie étale)* ; si k = C, la cohomologie étale à valeur dans $\mathbf{Z}/n\mathbf{Z}$ coïncide avec la cohomologie transcendante. À partir des mêmes idées, on peut même développer une théorie de l'homotopie des variétés algébriques (M. Artin).

7. Intersections

Définitions

Soit Y un fermé irréductible d'une variété algébrique X. La *codimension* de Y dans X est définie comme la borne supérieure des entiers n tels qu'il existe une suite strictement croissante $(F_0, F_1, \ldots, F_n)$ de fermés irréductibles de X avec F_0 = Y ; si X est irréductible :

$$\text{codim}_X(Y) = \dim X - \dim Y.$$

Supposons maintenant que X est une variété sans point singulier, et considérons des fermés irréductibles Y et Z de X ; si W est une composante irréductible de Y ∩ Z, on a l'inégalité suivante :

$$\text{codim } W \leqslant \text{codim } Y + \text{codim } Z$$

(par exemple, dans l'intersection de deux hypersurfaces, toutes les composantes irréductibles sont de codimension au plus 2). On dit que Y et Z se coupent *proprement* en W dans le cas où il y a égalité :

$$\text{codim } W = \text{codim } Y + \text{codim } Z$$

(formule des dimensions) ; on peut alors définir un entier m_W appelé *multiplicité* de W dans l'intersection de Y et Z (cf. **COURBES ALGÉBRIQUES,** pour le cas où Y et Z sont des courbes). Si Y et Z se coupent proprement en toutes les composantes irréductibles de $Y \cap Z$, on dit que l'intersection de Y et Z est propre, et on peut définir un « cycle » intersection :

$$Y \, . \, Z = \sum m_W W$$

(somme étendue aux composantes irréductibles de $Y \cap Z$). D'une manière générale, nous appellerons cycle de codimension m une combinaison linéaire formelle à coefficients entiers de fermés irréductibles de codimension m, et nous désignerons par $Z^m(X)$ le groupe additif des cycles de codimension m. On peut aussi définir la notion de cycles qui se coupent proprement ; si $z \in Z^m(X)$ et $z' \in Z^{m'}(X)$ sont des cycles qui se coupent proprement, leur *produit d'intersection* $z.z'$ est défini, et c'est un cycle de codimension $m + m'$. Voici les propriétés fondamentales du produit d'intersection :

(1) Soit $z \in Z^m(X)$ et $z'_1, z'_2 \in Z^{m'}(X)$; si $z.z'_1$ et $z.z'_2$ sont définis, il en est de même de $z.(z'_1 + z'_2)$ et $z.(z'_1 + z'_2) = z.z'_1 + z.z'_2$.

(2) Commutativité : $z.z' = z'.z$ (si l'un des membres est défini, l'autre l'est aussi).

(3) Associativité : $z.(z'.z'') = (z.z').z''$ lorsque les deux membres sont définis.

Soit $z \in Z^m(X)$ et $t \in Z^n(Y)$ où X et Y sont des variétés sans singularité ; on définit un cycle produit $z \times t \in Z^{m+n}(X \times Y)$.

Alors si $z.z'$ et $t.t'$ sont définis, il en est de même de $(z \times t)(z' \times t')$ et donc $(z \times t)(z' \times t') = (z.z') \times (t.t')$.

Soit $f : X \to Y$ un morphisme de variétés irréductibles sans singularité. On définit *l'image directe* f_* : $Z^m(X) \to Z^{m+r}(Y)$ $(r = \dim Y - \dim X)$; c'est une application additive telle que :

$$f_*(W) = 0 \text{ si } \dim f(W) < \dim W,$$
$$f_*(W) = d\,f(W) \text{ si } \dim f(W) = \dim W,$$

avec $d = [K(W) : K(f(W))]$ (degré de W comme revêtement ramifié de f(W) ; W est un fermé irréductible de X). *L'image réciproque* $f^*(t)$ d'un cycle t sur Y n'est pas toujours définie ; on pose $f^*(t) = p_*((X \times t).\Gamma_f)$ où p est la projection de $X \times Y$ sur X et Γ_f est le graphe de f. La codimension est conservée par f^*, et on a $f^*(t_1 + t_2) = f^*(t_1) + f^*(t_2)$ et $f^*(t.t') = f^*(t).f^*(t')$ lorsque les deux membres sont définis. Si des cycles z et z' sur X se coupent proprement, on a $z.z' = \delta^*_X(z \times z')$ où $\delta_X : x \mapsto (x, x)$ est le morphisme diagonal de X dans $X \times X$. Considérons un cycle z sur X et un cycle t sur Y ; on a la *formule de projection* $f_*(z . f^*(t)) = f_*(z) \, t$ (si les deux membres sont définis).

La théorie des intersections utilise diverses notions d'équivalence de cycles (linéaire, algébrique, numérique ; cf. **COURBES ALGÉBRIQUES)** ; ces relations d'équivalence sont compatibles avec l'addition des cycles et avec le produit d'intersection (et même avec les opérations f_* et f^*). Si z et z' sont des cycles quelconques sur une variété X, le produit $z.z'$ n'est pas défini en général, mais il existe un cycle z'_1 équivalent à z' qui coupe proprement z ; il en

résulte une loi quotient partout définie. Les propriétés (1), (2) et (3) montrent que l'ensemble des classes de cycles a une structure *d'anneau commutatif.*

Théorème de Riemann-Roch

Soit 3 un faisceau cohérent sur une variété algébrique projective X sans singularité. La *caractéristique d'Euler-Poincaré* de 3 est définie par :

$$\chi(X, \mathcal{F}) = \Sigma(-1)^i \dim H^i(X, \mathcal{F}).$$

Le théorème de Riemann-Roch exprime $\chi(X, 3)$ au moyen de classes de cycles liées à 3 et à X jouant le rôle de classes de Chern. Par exemple, si $\mathfrak{L}$ est un faisceau localement libre de rang 1, il s'interprète comme le faisceau des sections d'un fibré linéaire $\mathfrak{L}$; si s est une section rationnelle de $\mathfrak{L}$, on lui associe un diviseur $(s) = (s)_0 - (s)_\infty$, c'est-à-dire un cycle de codimension 1 sur X. On désigne par $(s)_0$ l'image réciproque pars de la section nulle de C ; $(s)_\infty$, se construit de même à l'aide de $1/s$. Lorsque X C P,(k) et que $\mathfrak{L} = \mathcal{O}_X(1)$ est le faisceau fondamental, (s) est l'intersection de X avec un hyperplan de P,(k). Lorsque s varie, le diviseur (s) reste dans une même classe pour l'équivalence linéaire, et cette classe D caractérise $\mathfrak{L}$ à isomorphisme près (première classe de Chern).

Si X est une courbe, le théorème de Riemann-Roch donne, pour un faisceau $\mathfrak{L}$ localement libre de rang 1 :

$$\chi(X, C) = \dim H^0(X, \mathfrak{L}) - \dim H^1(X, \mathfrak{L}) = \deg D + 1 - g$$

où deg D est le degré de la classe D, c'est-à-dire la somme des coefficients d'un diviseur quelconque appartenant à cette classe, et g est le genre de X (cf. COURBES ALGÉBRIQUES).

Supposons maintenant que X est une surface ; la formule de Riemann-Roch s'écrit :

$$\chi(X, \mathfrak{L}) = \dim H^0(X, \mathfrak{L}) - \dim H^1(X, C) + \dim H^2(X, \mathfrak{L}) = \frac{1}{2} \deg D.(D - K) + 1 + p_a,$$

où K est une classe de diviseurs, dite canonique, et liée au faisceau des formes différentielles sur X (ou au fibré tangent à X) et où p_a est un entier appelé *genre arithmétique* de X. Les invariants numériques, comme le genre arithmétique, ont été introduits initialement dans l'espoir d'arriver à une classification des variétés algébriques.

8. Groupes algébriques

On appelle groupe algébrique une variété algébrique G munie d'un morphisme m : G X G ⟶ G tel que pour toute variété algébrique T, l'application $(u, v) \mapsto m \circ (u, v)$ soit une loi de groupe sur l'ensemble G(T) des morphismes de T dans G ; si T est une variété affine, d'algèbre A, on écrit souvent G(A) au lieu de G(T) ; par exemple si T est la variété réduite à un point avec l'algèbre k, G(T) = $G(k)$ s'identifie à l'ensemble des points de G (cf. chap. 2), et on voit que m définit sur cet ensemble une structure de groupe. La théorie des groupes algébriques est assez analogue à celle des groupes de Lie, mais ses méthodes sont différentes.

Comme premiers exemples de groupes algébriques, citons le *groupe additif* $\mathbf{G}_a$, c'est-à-dire la droite affine $(k, k[t])$ munie de l'addition comme loi, et le *groupe multiplicatif* $\mathbf{G}_m$, c'est-à-dire la variété affine $(k - \{0\}, k\ [t, 1/t])$ munie de la multiplication ; on démontre que tout groupe algébrique affine de dimension 1

qui est connexe et réduit est isomorphe à G, ou à G,. L'ensemble $GL(n,k)$ des matrices carrées inversibles d'ordre n est un ouvert affine dans M(n, k) C $K^{n\times n}$, défini par det $(u_{ij}) \neq 0$; la multiplication des matrices en fait un groupe algébrique affine. Tout groupe algébrique affine s'identifie à un sous-groupe fermé d'un $GL(n, k)$; les groupes classiques sont des exemples de groupes algébriques (cf. GROUPES Groupes classiques et géométrie).

On peut également définir la notion d'opérations algébriques d'un groupe algébrique sur une variété algébrique. La théorie des groupes algébriques affines, édifiée par A. Borel, repose sur le théorème suivant :

Considérons un groupe algébrique affine G *résoluble* et *connexe*, qui opère sur une variété *complète* X. Il existe un point de X invariant par les opérations de G.

En appliquant ce résultat à un sous-groupe algébrique résoluble et connexe de $GL(n, k)$ opérant sur la « variété des drapeaux » de k^n, on trouve qu'un tel sous-groupe est conjugué d'un sous-groupe formé de matrices triangulaires (*théorème de Lie-Kolchin).* On appelle *sous-groupe de Bore1* d'un groupe algébrique affine G tout sous-groupe fermé résoluble connexe maximal ; les sous-groupes de Bore1 sont conjugués par automorphismes intérieurs. Si B est un sous-groupe de Bore1 de G, l'espace homogène G/B est une variété algébrique projective ; les sous-groupes H qui contiennent un sous-groupe de Bore1 (sous-groupes paraboliques) sont caractérisés par le fait que G/H est une variété complète.

À l'opposé des groupes affines se trouvent les *variétés abéliennes,* c'est-à-dire les groupes algébriques connexes réduits qui sont complets en tant que variétés algébriques. Si A est une variété abélienne, c'est une variété projective et sa loi de groupe est commutative. Lorsque le corps de base k est celui des nombres complexes, l'espace analytique A"" associé à une variété abélienne A est un tore complexe C^n/Γ (Γ sous-groupe additif discret de rang 2 n) ; on peut caractériser les tores complexes qui proviennent d'une variété abélienne par l'existence d'une *forme de Riemann* (c'est-à-dire une forme hermitienne positive non dégénérée sur C^n X C^n dont la partie imaginaire prend des valeurs entières sur Γ X Γ). Tout groupe algébrique connexe réduit G contient un sous-groupe affine connexe distingué H tel que le quotient G/H soit une variété abélienne.

À une variété algébrique complète X on associe des variétés abéliennes intéressantes : la *variété de Picard* et la *variété d'Albanese.* La première a pour ensemble sous-jacent l'ensemble des classes pour l'équivalence linéaire de diviseurs algébriquement équivalents à 0 sur X ; la seconde, A, est munie d'un morphisme X $\longrightarrow$ A tel que tout morphisme de X dans une variété abélienne « se prolonge » à A d'une manière unique. Si X est une courbe, la variété de Picard et la variété d'Albanese coïncident, et portent le nom de *jacobienne* de X ; la dimension de la jacobienne est égale au genre de X.

CHRISTIAN HOUZEL

Bibliographie

J. BOCHNAK, M. COSTE & M. F. ROY, *Géométrie algébrique* réelle. Springer-Verlag, 1987 / A. BOREL, *Linear Algebraic Groups*, Springer-Verlag. New York, 2' éd. 1991 / J. DIEUDONNÉ, *History* of *Algebraic Geometry*, Brooks/Cole Publ., Pacific Grove (Calif.),1985; Cours *de géométrie algébrique*, 2 vol., P.U.F.. Paris. 1974 / J. DIEUDONNÉ & A. GROTHENDIECK, Éléments *de géométrie algébrique,* 4 vol., Paris, 1960-1965 / A. GROTHENDIECK. *Fondements de la géométrie algébrique*, 2 vol.,

Secrétariat mathématique, Paris, 1985 / P. GRIFFITHS & J. HARRIS, *Principles of Algebraic Geometry*, J. Wiley, New York, 1978 / J. HARRIS, *Algebraic Geometry : a First Course*, Springer-Verlag, New York, 1992 / R. HARTSHORNE, *Algebraic Geometry*, *ibid.*, 1991 / C. A. PARIKH, *The Unreal Life of Oscar Zariski*, Academic Press, San Diego (Calif.), 1990 / I. R. ŠAFAREVIČ, *Basic Algebraic Geometry*, trad. du russe par H. A. Hirsh, Springer, Berlin-New York, 1990 / A. WEIL, *Foundations of Algebraic Geometry*, American Mathematical Society, Providence (R. I.).

GÉOMÉTRIE DIFFÉRENTIELLE CLASSIQUE

L'histoire des courbes planes est intimement liée à l'histoire et aux développements du calcul infinitésimal, et les premiers résultats obtenus au XVIIe siècle sont directement issus de considérations géométriques et cinématiques. Les courbes dans l'espace à trois dimensions (dites à « double courbure ») ont été étudiées par Clairaut (1731). C'est Monge qui, dans un mémoire présenté en 1771, introduisit les notions fondamentales de rayon de courbure et de surface réglée développable engendrée par les tangentes à une courbe gauche. En 1826, Cauchy définit la normale principale et donna des expressions de la courbure et de la torsion. Enfin, Frénet (1847) et Serret (1850) démontrèrent l'équivalent des formules qui portent leur nom.

Les surfaces, pour leur part, ont été au XVIIIe siècle une occasion naturelle de développer les fonctions de plusieurs variables. Euler, Monge admettent implicitement l'existence du plan tangent, qui est établie par Dupin en 1813 et reprise par Cauchy (1826). L'étude de la courbure entreprise par Euler (1760), qui introduisit les directions principales, a été approfondie par Meusnier (1776), Monge (1784) qui introduisit les lignes de courbure et Dupin (1813) qui introduisit les directions conjuguées et l'indicatrice qui porte son nom.

La contribution fondamentale de Gauss (*Disquisitiones circa superficies curvas*, 1827) donna un nouveau visage à la géométrie différentielle. Il utilisa une représentation paramétrique des surfaces (amorce de la notion de carte locale) et dégagea le caractère intrinsèque de la courbure totale ; tous les résultats du chapitre 6 lui sont dus. Enfin, le tournant décisif est dû à Riemann (*Sur les hypothèses qui servent de fondement à la géométrie*, 1854) qui surmonta les difficultés rencontrées pour donner une définition globale mathématiquement satisfaisante des courbes et des surfaces en introduisant la notion de variété à n dimensions.

On se bornera dans cet article à l'étude des courbes et surfaces plongées dans l'espace euclidien à deux ou trois dimensions. Les définitions correctes exigent l'emploi du théorème des fonctions implicites, c'est pourquoi on introduira d'abord ici la notion d'arc paramétré, dont l'image est une trajectoire, puis on effectuera l'étude locale ; la notion de courbe s'obtiendra ensuite en « recollant » de manière régulière une réunion de trajectoires. On définira de même les surfaces en recollant entre elles des images de représentations paramétriques régulières ; on introduira les deux formes fondamentales (l'ensemble de ces deux formes définit localement la surface à un déplacement euclidien près) et la courbure totale (qui ne dépend que de la première forme fondamentale). Cette courbure totale joue un grand rôle, tant dans l'étude locale (position par rapport au plan tangent) que globale (caractéristique d'Euler-Poincaré) des surfaces.

1. Sur quelques propriétés de l'espace euclidien

La structure E_3, *d'espace euclidien* de $\mathbf{R}^3$ est définie par le choix du produit scalaire usuel pour lequel la base canonique $\varepsilon_1 = (1, 0, 0)$, $\varepsilon_2 = (0, 1, 0)$, $\varepsilon_3 = (0, 0, 1)$ est orthonormée ; la norme de X = (x, y, z) est alors :

$$x = \sqrt{x^2 + y^* + z^2}.$$

Un *déplacement euclidien* D est une application affine de $\mathbf{R}^3$ dans $\mathbf{R}^3$ telle que l'application linéaire associée soit une *rotation*, c'est-à-dire une transformation orthogonale de déterminant égal à 1 ; l'ensemble des déplacements euclidiens forme alors un groupe $\mathcal{G}$, produit semi-direct du groupe additif $\mathbf{R}^3$ (groupe des translations) et du groupe $\mathbf{O}^+(3, \mathrm{R})$ des rotations. Un déplacement euclidien est défini par les trois composantes (a', a^2, a') de la translation et les éléments a_j^i d'une matrice orthogonale de déterminant 1. On appelle *mouvement euclidien* une application $t \mapsto$ D, d'un intervalle I de R dans le groupe $\mathcal{G}$; ce mouvement sera dit différentiable de classe C^k si les fonctions $t \mapsto a^i(t)$ et $t \mapsto a_j^i(t)$ définissant ce mouvement sont des fonctions k fois continûment dérivables de t. Dans une interprétation cinématique, le paramètre t désigne le temps. Dans ce qui suit, nous adopterons le langage de la cinématique (vitesse, accélération) pour un paramétrage quelconque.

Un *repère* $\mathcal{T}$ (M, e_1, e_2, e_3) de l'espace euclidien E_3 est le transformé par un déplacement D du repère canonique (0, ε_1, ε_2, ε_3) ; (e_1, e_2, e_3) est donc une base orthonormée de E_3 (considéré comme espace vectoriel) de même sens que (ε_1, ε_2, ε_3), ce qui *oriente* l'espace. De plus, étant donné deux repères $\mathcal{T}$ et $\mathcal{T}'$, il existe un déplacement euclidien et un seul transformant $\mathcal{T}$ en $\mathcal{T}'$.

À un mouvement euclidien $t \mapsto D_t$ correspond un *repère mobile* $\mathcal{T}_t = D_t(\mathcal{T}_0)$ (on peut également définir un repère dépendant de plusieurs paramètres) défini par :

$$M(t) = a^1(t)\varepsilon_1 + a^2(t)\varepsilon_2 + a^3(t)\varepsilon_3$$
$$e_i(t) = a_i^1(t)\varepsilon_1 + a_i^2(t)\varepsilon_2 + a_i^3(t)\varepsilon_3$$

pour i = 1, 2, 3 ; si on rapporte les vecteurs dérivés $d\mathrm{M}/dt$ et de_i/dt au repère mobile, on a, en utilisant le fait que $d(e, e_j)/dt = 0$:

$$\frac{d\mathrm{M}}{dt} = \alpha^1 e_1 + \alpha^2 e_2 + \alpha^3 e_3,$$
$$\frac{de_1}{dt} = re_2 - qe_3,$$
$$\frac{de_2}{dt} = -re_1 + pe_3,$$
$$\frac{de_3}{dt} = qe_1 - pe_2.$$

Le vecteur ω de composantes (p, q, r) est le vecteur *rotation instantanée du mouvement* ; pour tout vecteur V lié à $\mathcal{T}_t$, c'est-à-dire tel que :

$$\mathrm{V} = ne, + be_2 + ce_3,$$

où a, b, c sont des constantes, on a alors :

$$\frac{d\mathrm{V}}{dt} = \omega \wedge \mathrm{V}.$$

On démontre que **toute** application continue d'une partie de E_3 dans E_3, qui est une isométrie (c'est-à-dire conserve la norme), est une restriction de déplacement euclidien ou d'antidéplacement (transformation affine dont la transformation linéaire associée est orthogonale de déterminant 1).

Rappelons enfin (cf. CALCUL INFINITÉSIMAL Calcul à plusieurs variables) qu'une applicationfd'un ouvert **U** de $\mathbf{R}^p$ dans $\mathbf{R}^3$ est dite différentiable de classe C^k si f

s'exprime au moyen de trois fonctions numériques. $f = (f^1, f^2, f^3)$, admettant des dérivées partielles continues jusqu'à l'ordre k. La *différentielle* $Df(a)$ de f au point $a = (a^1, a^2, \ldots, a^p) \in U$ est l'application linéaire de $\mathbf{R}^p$ dans $\mathbf{R}^3$ définie par :

$$Df(a) \, . \, h = \sum_{i=1}^{p} \frac{\partial f}{\partial u^i}(a) h^i,$$

pour $h = (h^1, \ldots, h^p) \in \mathbf{R}^p$; on considérera souvent la forme quadratique associée à la différentielle seconde en a, définie par :

$$D^2 f(a) \, . \, (h, h) = \sum_{i,j=1}^{p} \frac{\partial^2 f}{\partial u^i \partial u^j}(a) h^i h^j ;$$

on appelle *application affine tangente* en a à f l'application affine $T_a f$ définie par :

$$T_a f : h \mapsto f(a) + Df(a) \, . \, h,$$

ce qui équivaut à :

$$f(a + h) = T_a f(h) + \varepsilon(h) \| h \|,$$

où $\varepsilon(h)$ tend vers 0 quand h tend vers 0.

Rappelons enfin que, si on compose deux applications différentielles g et f, on a :

$$D(f \circ g)(b) = Df(g(a)) \circ Dg(b)$$

et

$$T_b(f \circ g) = T_{g(b)} f \circ T_b g ;$$

en particulier, si $Dg(b)$ (et par suite $T_b g$) est une bijection, alors $T_b(f \circ g)$ et $T_{g(b)} f$ ont la même image. Dans le cas particulier où g est une fonction d'une variable, alors on a :

$$\frac{d(f \circ g)}{dt}(t_0) = Df(g(t_0)) \, . \, \frac{dg}{dt}(t_0).$$

2. Remarques sur les courbes et les surfaces

On a une notion intuitive de « courbe » dans l'espace euclidien à 2 ou 3 dimensions : une courbe de E_2 est définie par une équation $F(x, y) = 0$, ou $y = f(x)$; une courbe de E_3, est définie par deux équations $z = g(x)$ et $y = f(x)$, ou $F(x, y, z) = 0$ et $G(x, y, z) = 0$. De même, une « surface » de E_3 est définie par une équation $z = f(x, y)$, ou $F(x, y, z) = 0$.

Mais, si on veut préciser ces notions, des difficultés surgissent. Par exemple, le cercle de centre 0 et de rayon R est l'ensemble des points de E_2 dont les coordonnées vérifient l'équation :

$$x^2 + y^2 = R^2,$$

mais on peut aussi représenter ce cercle par :

$$\begin{cases} x = R \cos t \\ y = R \sin t, \; 0 \leqslant t \leqslant 2\pi ; \end{cases}$$

or l'application ainsi définie de l'intervalle fermé $[0, 2\pi]$ sur le cercle n'est pas biunivoque, car les extrémités de cet intervalle sont appliquées sur le même point A (1, 0) du cercle. Or, ce point ne présente aucune singularité sur le cercle.

De même, la sphère de centre 0 et de rayon R est l'ensemble des points de E_3 dont les coordonnées vérifient :

$$x^2 + y^2 + z^2 = R^2 ;$$

mais elle peut aussi être représentée par :

$$\begin{cases} x = R \cos u \cos t \\ y = R \cos u \sin t \\ z = R \sin u \end{cases}$$

pour $0 \leqslant t \leqslant 2\pi$ et $\pi/2 \leqslant u \leqslant \pi/2$, les courbes $u =$ constante étant les parallèles et les courbes $t =$ constante étant les méridiens Mais l'application ainsi définie d'un rectangle de E_2 sur la sphère n'est pas biunivoque : les « pôles » P (0, 0, 1) et P' (0, 0, $-$ 1) correspondent respectivement à $u = \pi/2$ et $u = -\pi/2$, t quelconque ; pourtant les points P et P' ne présentent aucune singularité sur la sphère.

Par contre, le cône de révolution d'axe Oz d'équation :

$$x^2 + y^2 - z^2 = 0, \quad z \geqslant 0$$

est en correspondance bijective avec le plan d'équation z = 0, cette correspondance associant au point m (x, y) le point M(x, y, z) tel que :

$$z = \sqrt{x^2 + y^2};$$

pourtant ce cône présente un point singulier qui est son sommet.

3. Arcs paramétrés et trajectoires

Nous allons distinguer à présent les notions d'arc paramétré et de courbe régulière.

On appellera *arcparamétré* de classe C^k une applicationfd'un intervalle 1 = [a, b] de R dans E_2 ou E_3 qui soit k fois continûment dérivable dans 1 (en a et b, on considère respectivement les dérivées à droite et à gauche) ; dans ce qui suit, on supposera k assez grand pour que toutes les dérivations effectuées aient un sens. On appelle *trajectoire* de l'arc paramétré f le sous-ensemble image A $= f(I)$.

On dira que deux arcs paramétrés (f, I) et (g, J) de classe C^k sont *C^k-équivalents* s'il existe un C^k-difféomorphisme φ de I sur J (c'est-à-dire une bijection k fois continûment dérivable ainsi que son inverse) tel que :

$$f = g \circ \varphi;$$

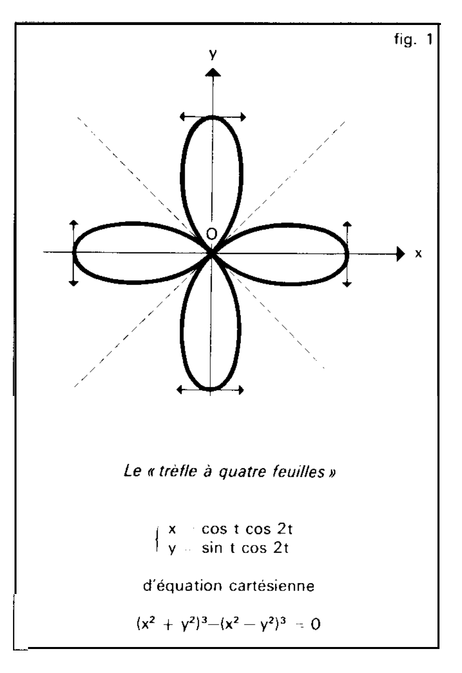
cela entraîne en particulier que les deux arcs ont la même trajectoire. On dira que le changement de loi de « temps » $\tau = \varphi(t)$ est un changement de paramètre *admissible*. Pour **tout** $t \in 1$, on a $d\varphi/dt \neq 0$; les paramètres admissibles se répartissent donc en deux classes : ceux pour lesquels $d\varphi/dt > 0$ et ceux pour lesquels $d\varphi/dt < 0$. Choisir un signe revient à *orienter* la trajectoire.

Exemples

Considérons le « trèfle à quatre feuilles » :

$$x = \cos t\,(\cos^2 t - \sin^2 t) = \cos t \cos 2t$$
$$y = \sin t\,(\cos^2 t - \sin^2 t) = \sin t \cos 2t$$

pour $0 \leqslant t \leqslant 2\pi$; l'origine 0 est un *point multiple* pour la trajectoire, car on a $x = y = 0$ pour $t = \pi/4$, $3\pi/4$, $5\pi/4$, $7\pi/4$ (fig. 1).



Le « trèfle à quatre feuilles »

$\begin{cases} x = \cos t \cos 2t \\ y = \sin t \cos 2t \end{cases}$

d'équation cartésienne

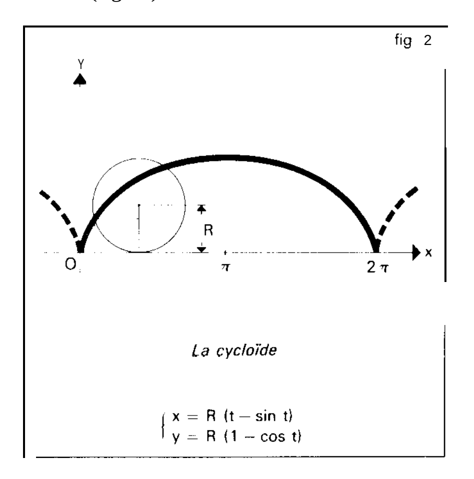
$(x^2 + y^2)^3 - (x^2 - y^2)^3 = 0$

Soit maintenant, pour $t \in R$, l'arc paramétré :

$$x = R(t - \sin t),$$
$$y = R(1 - \cos t);$$

la trajectoire est composée d'arcs se déduisant les uns des autres par translation.

C'est la *cycloïde*, trajectoire d'un point lié à un cercle qui roule sans glisser sur une droite (fig. 2).



La cycloïde

$$\begin{cases} x = R\,(t - \sin t) \\ y = R\,(1 - \cos t) \end{cases}$$

Points réguliers

Soit f un arc paramétré. On appelle *vitesse* à l'instant *t* le vecteur dérivé :

$$\frac{df}{dt}(t) = \lim_{h \to 0} \frac{f(t+h) - f(t)}{h},$$

si on change la loi de temps, $t = \varphi(\tau)$ et $g(\tau) = f(\varphi(\tau))$, on a :

$$\frac{dg}{d\tau}(\tau) = \frac{df}{dt}(t)\frac{\varphi}{d\tau}(\tau)$$

et les vecteurs $(df/dt)(t)$ et $(dg/d\tau)(\tau)$, par suite, sont colinéaires **ou** simultanément tous deux nuls.

Si $(df/dt)(t)$ n'est pas nui et si t est un point intérieur à I, la droite portant le vecteur vitesse s'appelle la *tangente* en M à la trajectoire. Si $I = [a, b]$, on dit que l'arc est fermé lorsque $f(a) = f(b)$; remarquons que, même si en tout point la dérivée est non nulle, la trajectoire peut ne pas avoir de tangente en $f(a)$ (par exemple si $x = \cos t \cos 2t$, $y = \sin t \cos 2t$ pour $\pi/4 \leqslant t \leqslant 3\pi/4$).

Si t est intérieur à l'intervalle I et si $f'(t) \neq 0$, on dit que le point $f(t)$ est un point *régulier* de la trajectoire ; cette propriété se conserve par changement de paramètre admissible. Il résulte alors du théorème des fonctions implicites (cf. CALCUL INFINITÉSIMAL Calcul à plusieurs variables) qu'il existe un intervalle $I_1 \subset I$ contenant t tel que la restriction de f à I_1, soit injective : par exemple, tous les points du cercle $x = R \cos t$, $y = R \sin t$ sont des points réguliers et la restriction de f à

$$]t - \pi, t + \pi[\cap I$$

est injective pour tout t.

Pour un arc paramétré, le vecteur d^2f/dt^2 représente l'accélération. Dans un changement de paramètre admissible, on a, pour $g(\tau) = f(t)$:

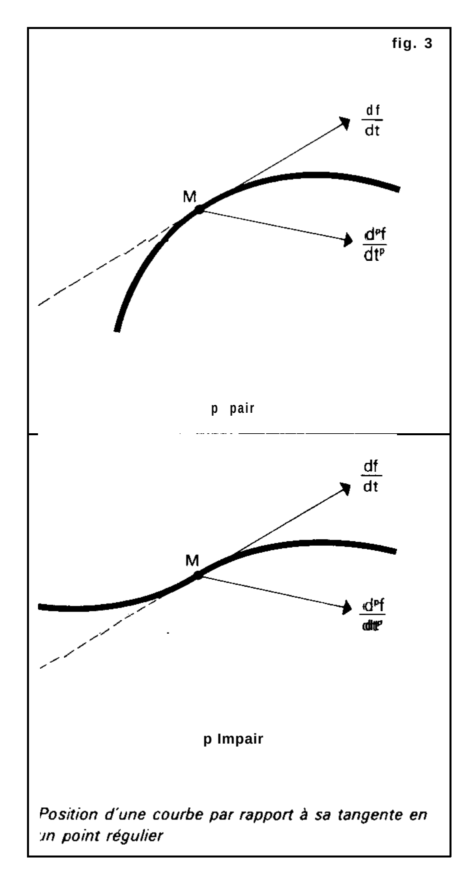
$$\frac{d^2g}{d\tau^2} = \left(\frac{dt}{d\tau}\right)^2 \frac{d^2f}{dt^2} + \frac{d^2t}{d\tau^2}\frac{df}{dt};$$

ainsi, si les vecteurs df/dt et d^2f/dt^2 ne sont pas colinéaires, le vecteur $d^2g/d\tau^2$ appartient au plan engendré par ces vecteurs, appelé *plan osculateur* à la trajectoire **au** point $f(t)$.

En un point régulier, désignons par p le plus petit entier $\geqslant 2$ tel que le vecteur d^pf/dt^p soit non nul et non colinéaire au vecteur vitesse df/dt : au voisinage d'un tel point, la trajectoire présente l'aspect indiqué sur la figure 3 : le point est dit ordinaire si p est pair ; si p est impair, la trajectoire « traverse » la tangente au voisinage du point et on dit qu'il y a *inflexion*. Le cas où df/dt et d^2f/dt^2 sont colinéaires se ramène au précédent. car on peut trouver. dans un intervalle $I_1 \subset I$ contenant t, un changement de paramètre admissible tel que $d^2g/d\tau^2$ soit nul.

Points singuliers

Soit maintenant t_0 une valeur du paramètre pour laquelle le vecteur vitesse est nul ; on



Position d'une courbe par rapport à sa tangente en un point régulier

dit que le point $f(t_0)$ est un *point singulier.* Soit p et q, $p < q$, les plus petits entiers tels que les vecteurs :

$$\frac{d^pf}{dt^p}(t_0) \quad \text{et} \quad \frac{d^qf}{dt^q}(t_0)$$

soient tous deux non nuls et non colinéaires.

Pour p *impair*, posons $\tau = (t - t_0)^p$; on définit ainsi un changement de paramètre qui est un homéomorphisme, mais qui n'est pas admissible au sens ci-dessus, car la fonction réciproque $\tau \mapsto t = t_0 + \tau^{1/p}$ n'est pas dérivable pour $\tau = 0$. On a alors :

$$\frac{dg}{d\tau}(0) = \frac{1}{p!}\frac{d^pf}{dt^p}(t_0),$$

et, par suite, l'arc paramétré défini par l'application g est régulier pour $\tau = 0$ et il a la même trajectoire que l'arc défini par f. On peut dire qu'on a une singularité « cinématique », due au paramétrage, et non une singularité de la trajectoire. Par exemple, l'arc paramétré défini par $x = t^3$, $y = t^6$, $z = t^9$ a une singularité pour $t = 0$, mais, si on pose $\tau = t^3$, on obtient $x = \tau$, $y = \tau^2$, $z = \tau^3$, qui est un arc paramétré sans singularité.

Si p est *pair,* on ne peut pas éliminer la singularité par changement de paramètre. On dit que la trajectoire présente un *point de rebroussement.* Au voisinage d'un tel point, la courbe présente l'aspect indiqué par la figure 4 ; si q est impair, on dit qu'on a un point de rebroussement de première espèce (cf. la cycloïde, fig. 2) ; si q est pair, on dit qu'on a un point de rebroussement de deuxième espèce.

Élément de longueur

Jusqu'à présent, nous n'avons pas utilisé la structure euclidienne de l'espace ; nous allons en faire usage pour choisir un paramètre admissible privilégié au voisinage d'un point régulier. On appelle *abscisse curviligne* un paramètres tel que, pour $g(s) = f(t)$, le vecteur vitesse dg/ds soit unitaire. Comme :

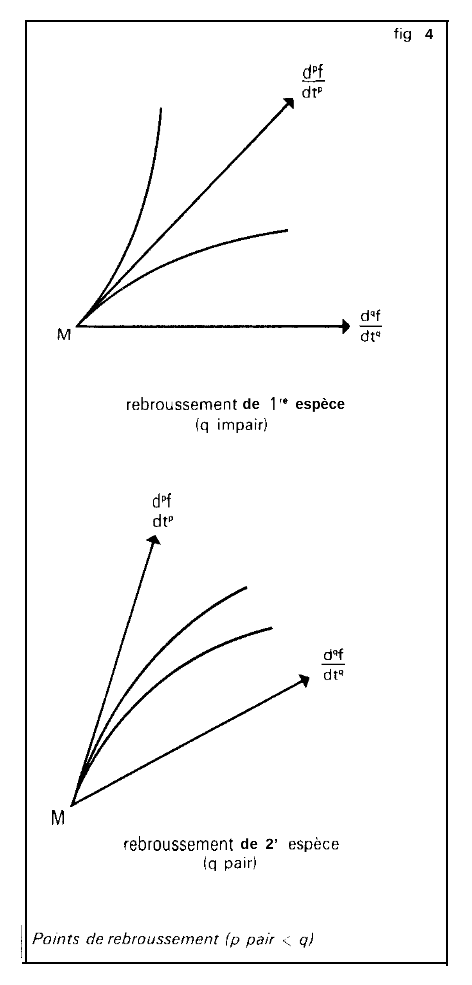
$$\left(\frac{dg}{ds}\right)^2 = \left(\frac{df}{dt}\right)^2\left(\frac{dt}{ds}\right)^2 = 1,$$

on doit avoir :

$$\left(\frac{ds}{dt}\right)^2 = \left(\frac{df}{dt}\right)^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2,$$

d'où :

$$s - s_0 = \pm\int_{t_0}^{t}\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\,dt,$$



Points de rebroussement (p pair < q)

choisir le signe de $s - s_0$, c'est *orienter* la trajectoire en choisissant un « sens de parcours ».

Si f définit un arc paramétré, l'intégrale :

$$\mathrm{L} = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt$$

s'appelle la longueur de l'arc $A = f([a, b])$. Cette longueur est invariante par changement de paramètre admissible et peut s'interpréter ainsi : à chaque subdivision finie $t_0 = a < t_1 < t_2 < \dots < t_n, \dots < t_n = b$ de l'intervalle $[a, b]$, associons la longueur de la ligne polygonale joignant les points $f(t_i)$:

$$\sum_{i=1}^{n} \|f(t_i) - f(t_{i-1})\|;$$

alors la longueur L est la borne supérieure de ces nombres pour toutes les subdivisions finies de $[a, b]$.

Par exemple, la longueur du cercle de rayon R est :

$$\int_0^{2\pi} \mathrm{R}\sqrt{\sin^2 t + \cos^2 t}\, dt = \mathrm{R}\int_0^{2\pi} dt = 2\pi\mathrm{R}.$$

Pour la cycloïde, la longueur d'un arc compris entre deux points de rebroussement successifs est :

$$s(2\pi) - s(0) = \int_0^{2\pi} 2\mathrm{R}\sin\frac{t}{2}\, dt = 8\mathrm{R}.$$

Trièdre de Frénet

Nous allons continuer l'étude locale d'un arc paramétré *au voisinage d'un point régulier* en associant à chaque point de la trajectoire un trièdre orthonormé de sens direct.

Si on prend pour paramètre au voisinage d'un point régulier l'abscisse curviligne s, alors, par définition, le vecteur vitesse :

$$\boldsymbol{t} = \frac{dg}{ds}$$

est unitaire. Si le vecteur dt/ds n'est pas nul, il est orthogonal au vecteur $\boldsymbol{t}$; on posera :

$$\frac{\boldsymbol{dt}}{ds} = \frac{n}{\mathrm{R}},$$

où $\boldsymbol{n}$ est un vecteur unitaire et R un nombre *positif* appelé rayon de courbure si la

courbe est non *plane* (dans le cas d'une courbe plane, on choisit le vecteur $\boldsymbol{n}$ *directement perpendiculaire* et R est un nombre réel de signe quelconque) ; on appelle alors centre de courbure le point P défini par $P = M + Rn$. Remarquons que si on change l'orientation de la trajectoire, $s \mapsto -s$ et $t \mapsto -t$, le vecteur dt/ds ne change pas d'orientation. Définissons alors le vecteur $\boldsymbol{b}$ par la formule :

$$\boldsymbol{b} = \boldsymbol{t} \wedge \boldsymbol{n};$$

le trièdre $\boldsymbol{t}$, $\boldsymbol{n}$, $\boldsymbol{b}$ s'appelle le trièdre de Frénet. On a les formules :

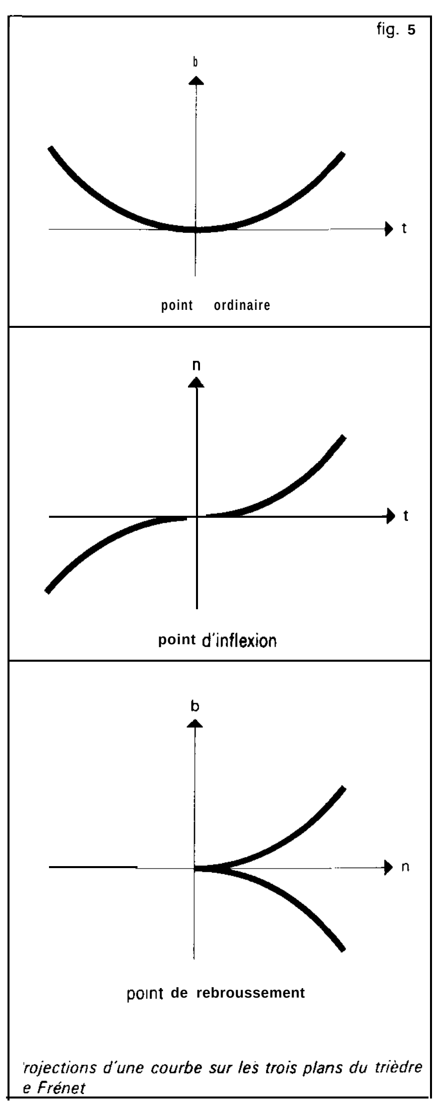
$$\frac{dt}{ds} \quad \frac{n}{R}$$

$$\frac{dn}{ds} = -\frac{t}{R} + \frac{b}{T},$$

$$\frac{db}{ds} \quad \frac{n}{T};$$

les fonctions $1/R(s)$ et $1/T(s)$ s'appellent respectivement la *courbure* et la *torsion* de la trajectoire. Au voisinage d'un point régulier, les projections de la trajectoire sur les trois plans définis par $\boldsymbol{t}$, $\boldsymbol{n}$, $\boldsymbol{b}$ présentent l'aspect indiqué par la figure 5.

On démontre que si deux trajectoires ont même courbure et même torsion pour tout s, alors il existe un déplacement euclidien transformant l'une en l'autre. Si $1/T = 0$ pour **tout** s, la trajectoire est plane ; si $1/R = 0$, alors la torsion est nulle et la trajectoire est une droite.



'rojections d'une courbe sur les trois plans du trièdre e Frénet

4. Courbes régulières

Nous sommes enfin capables de donner une définition correcte de la notion de courbe régulière.

Par définition, une *courbe régulière C*, de classe C^k, de l'espace euclidien E_3 ou E_2 est un sous-ensemble qui possède la propriété suivante : Tout point $x \in C$ est centre d'une boule ouverte B (resp. d'un disque ouvert B) telle qu'il existe un arc paramétréf: $1 \to E_3$ (ou E_2), de classe C^k, tel que $f'(t) \neq 0$ pour tout $t \in I$, qui soit un homéomorphisme de 1 sur $C \cap B$. Ainsi C est une réunion de trajectoires d'arc paramétrés sans points singuliers. Si (f_i, I_i)

et (f_j, I_j) sont deux représentations paramétriques telles que $I_{ij} = f_i(I_i) \cap f_j(I_j)$ ne soit pas vide, alors, $f_j^{-1} \circ f_i$, défini dans $f_i^{-1}(I_{ij})$, est un changement de paramètre admissible. Ce qui précède sur les arcs paramétrés montre qu'on peut définir la tangente en chaque point ainsi que, quand le plan osculateur est défini, le trièdre de Frénet. Remarquons qu'une courbe aussi simple que la cycloïde ne rentre cependant pas dans ce cadre, car elle présente des points singuliers.

Le théorème des fonctions implicites entraîne que si F est une fonction numérique sur E^2, différentiable de dérivée ne s'annulant pas sur $F^{-1}(a)$ pour un nombre réel a, l'ensemble $F^{-1}(a)$ est une courbe régulière ; de même, dans E_3, si deux fonctions F_1 et F_2 sont indépendantes, l'ensemble défini par F_1 = constante et F_2 = constante est une courbe régulière. Par exemple, dans le plan, l'équation :

$$(x^2 + y^2)^3 - (x^2 - y^2)^2 = a$$

représente une courbe régulière pour $a \neq 0$, car la différentielle de $F(x, y) = (x^2 + y^2)^3 - (x^2 - y^2)^2$ ne s'annule que pour $x = y = 0$, et alors $F(0, 0) = 0$. Par contre, l'équation $(x^2 + y^2)^3 - (x^2 - y^2)^2 = 0$ représente une courbe présentant une singularité à l'origine : c'est le « trèfle à quatre feuilles » vu ci-dessus (fig. 1).

Enfin. une courbe régulière est dite *orientée* si tous les changements de représentation paramétrique $f_j^{-1} \circ f_i$ sont des fonctions croissantes.

On peut démontrer pour les courbes *fermées* régulières planes les deux théorèmes suivants : l'angle dont « tourne » le vecteur unitaire tangent à une telle courbe orientée est $\pm 2\pi$; pour toute courbe fermée convexe (c'est-à-dire ne présentant pas d'inflexion, et par suite pour laquelle la courbure garde un signe constant) il y a au moins quatre *sommets* (c'est-à-dire des points où la courbure présente un extrémum).

5. Définition des surfaces

Surfaces régulières

On appellera *surface régulière* de classe C^k, $k \geqslant 1$, de l'espace euclidien E_3 un sous-ensemble $S \subset E_3$ possédant la propriété suivante : Tout point de S est centre d'une boule ouverte B de E_3 telle qu'il existe une application φ de classe C^k d'un ouvert U de R^2 dans E_3 :

$$(u, v) \mapsto \varphi(u, v)$$

de rang 2 en tout point de U, qui soit un homéomorphisme de U sur $S \cap B$. Si $\varphi = (\varphi_1, \varphi_2, \varphi_3)$, où φ_1, φ_2 et φ_3 sont des fonctions numériques de classe C^k, la condition sur le rang signifie que la matrice :

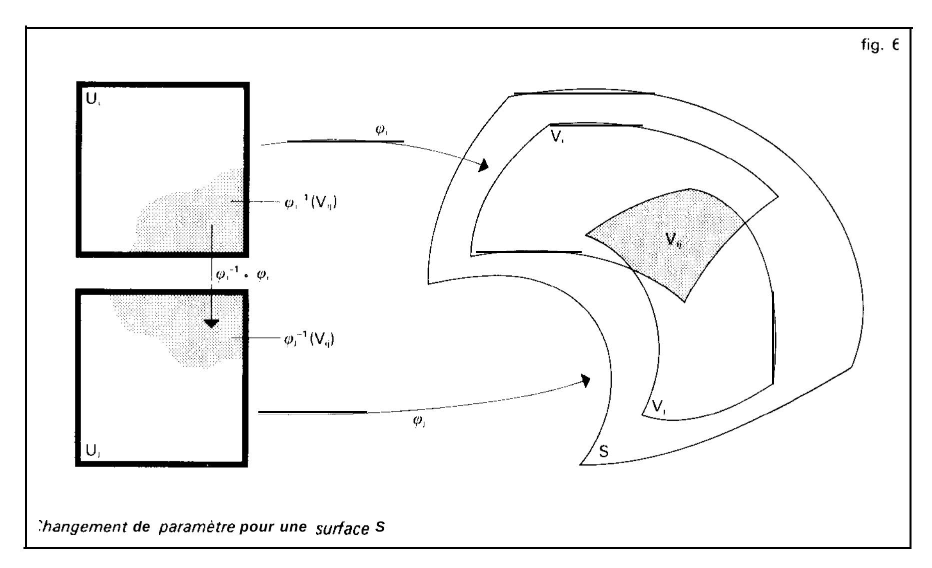
$$\begin{pmatrix} \frac{\partial \varphi_1}{\partial u} & \frac{\partial \varphi_2}{\partial u} & \frac{\partial \varphi_3}{\partial u} \\ \frac{\partial \varphi_1}{\partial v} & \frac{\partial \varphi_2}{\partial v} & \frac{\partial \varphi_3}{\partial v} \end{pmatrix}$$

est de rang 2, ou encore que le produit vectoriel :

$$\frac{\partial \varphi}{\partial u} \wedge \frac{\partial \varphi}{\partial v}$$

est un vecteur non nul.

L'application φ est appelée une *représentation paramétrique vraie*, ou *régulière*, de $V = S \cap B$. Il résulte alors du théorème des fonctions implicites qu'au voisinage de chaque point de S on peut exprimer l'une des coordonnées comme fonction de classe C^k des deux autres, l'application ainsi définie étant de rang 2. Si (φ_i, U_i) et



Changement de paramètre pour une surface S

(φ_j, U,) sont deux représentations paramétriques telles que $V_{ij} = \varphi_i(U_i) \cap \varphi_j(U_j)$ ne soit pas vide, le changement de paramètre $\varphi_j^{-1} \circ \varphi_i$ est un difféomorphisme de classe C^k de $\varphi_i^{-1}(V_{ij})$ sur $\varphi_j^{-1}(V_{ij})$ (fig. 6). Ces considérations conduisent directement à la notion de variété différentiable générale.

Par exemple, pour la sphère S de centre 0 et de rayon 1, la représentation paramétrique :

$$\begin{cases} x = \cos u \cos v \\ y = \cos u \sin v \\ z = \sin u \end{cases}$$

n'est pas réguliére aux pôles. Par contre, au moyen d'une projection stéréographique de pôle (P(0, 0, 1), on définit une application de classe C^k de $\mathbf{R}^2$ sur S — {P} :

$$x = \frac{2u}{1 + u^2 + v^2},$$
$$y = \frac{2v}{1 + u^2 + v^2},$$
$$z = \frac{u^2 + v^2 - 1}{1 + u^2 + v^2},$$

qui est un homéomorphisme. De même, l'inversion de pôle P'(0, 0, 1) applique $\mathbf{R}^2$ sur S — {P'} :

$$x = \frac{2u'}{1 + u'^2 + v'^2},$$
$$y = \frac{2v'}{1 + u'^2 + v'^2},$$
$$z = \frac{u'^2 + v'^2 - 1}{1 + u'^2 + v'^2},$$

dans $\mathbf{R}^2$ — {O}, le changement de paramètre est l'inversion de pôle 0 et de puissance 1 :

$$u' = \frac{u}{u^2 + v^2}, \quad v' = \frac{v}{u^2 + v^2}$$

En utilisant le fait que S est compacte (fermée et bornée dans $\mathbf{R}^3$), on peut montrer qu'il n'existe aucune représentation paramétrique régulière de la sphère tout entière.

De manière générale, soit f une fonction numérique de classe C^k définie dans E_3 ; si a $\in f(E_3)$ est tel que f soit de rang 1 (c'est-à-dire que sa différentielle ne s'annule pas) en tout point de f'(n), alors

le théorème des fonctions implicites entraîne que l'ensemble $f^{-1}(a)$ est une surface régulière. Par exemple, la sphère est définie par l'équation :

$$x^2 + y^2 + z^2 = 1$$

S'il existe des points de $f^{-1}(a)$ où la différentielle s'annule, on dit que $f^{-1}(a)$ est une *surface avec singularités* ; par exemple le cône $x^2 + y^2 \quad z^2 = 0$, vu au chapitre 2, a l'origine pour point singulier (sommet du cône).

Comme exemples importants de surfaces régulières, on a notamment les *quadriques* (à l'exclusion du cône) définies par une équation :

$$f(x, y, z) = \text{constante}$$

où f est un polynôme de degré 2 (cf. QUADRIQUES), par exemple *l'hyperboloïde à une nappe* :

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 - \left(\frac{z}{c}\right)^2 = 1;$$

il admet la représentation paramétrique :

$$\begin{cases} x = a \cos u \operatorname{ch} v \\ y = b \sin u \operatorname{ch} Y \\ z = c \operatorname{ch} v \end{cases}$$

qui n'est pas régulière, car non bijective. Par contre, cette représentation paramétrique définit un difféomorphisme de S, X R sur la surface, en désignant par S, le cercle de centre 0 et de rayon 1. Un autre exemple est le *paraboloïde hyperbolique* (« selle de cheval ») d'équation :

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 - 2z = 0,$$

qui est difféomorphe à $\mathbb{R}^2$ (fig. 10).

On dit qu'une surface S est *réglée* si par tout point de S passe au moins une droite entièrement contenue dans S ; une telle droite est appelée une *génératrice* de la surface (fig. 7 et 8) Par exemple, l'hyperboloïde à une nappe et le paraboloïde hyperbolique sont engendrés par deux familles à un paramètre de droites : par chaque point passe une génératrice de chaque famille.

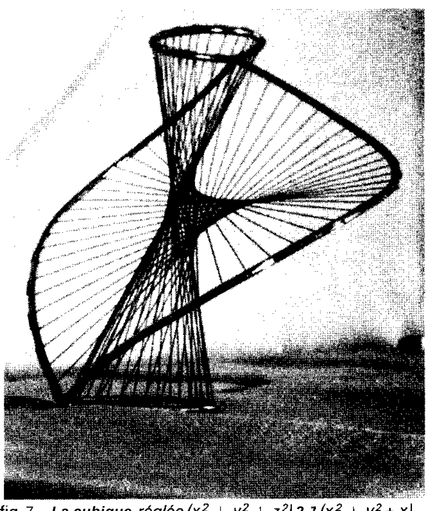


fig. 7 - *La cubique réglée $(x^2 + y^2 + z^2) 2 z (x^2 + y^2 + x) = 0$ (D. R.).*

Parmi les autres surfaces d'un type particulier, notons les *surfaces de révolution* : une surface S (régulière ou avec singularités) est dite de révolution autour d'un axe D si toute rotation d'axe D transforme S en elle-même. Ainsi, si M est un point de S qui n'appartient pas à D, le cercle d'axe D passant par M est entièrement situé dans S ; on dit qu'un tel cercle est un *parallèle* de la surface. De même, on appelle *méridien* les intersections de S avec les plans passant par D ; bien entendu, la terminologie précédente généralise celle adoptée traditionnellement pour la sphère qui est de révolution autour de tout axe passant par son centre. Par exemple, pour $a = b$,

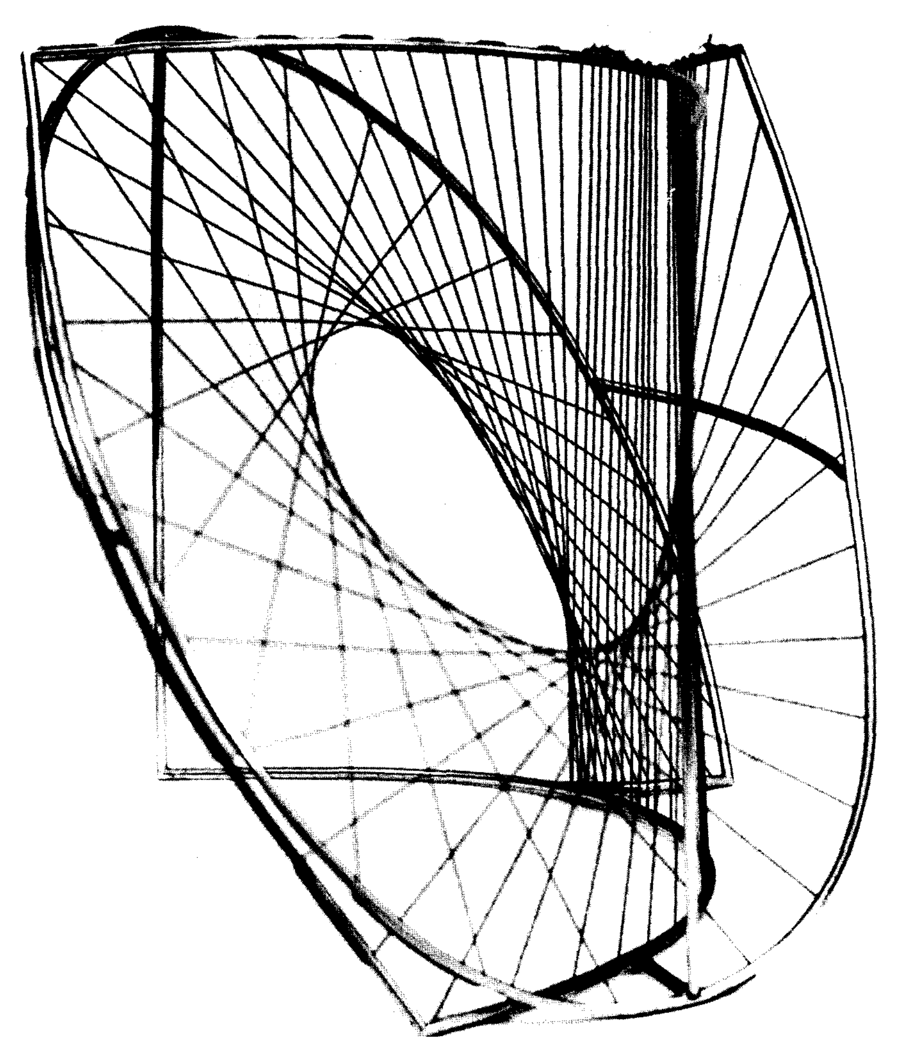


fig. 8 - *Une surface réglée du quatrième degré (M. Tristant).*

L'hyperboloïde à une nappe est de révolution autour de l'axe Oz ; c'est la surface engendrée par la rotation d'une droite autour d'un axe non coplanaire et les méridiens sont des hyperboles. Le *tore* est défini par la rotation d'un cercle autour d'une droite de son plan ne le rencontrant pas (fig. 11). Il admet la représentation paramétrique (non régulière, parce que non bijective) :

$$\begin{cases} x = (2 + \cos u)\cos v \\ y = (2 + \cos u)\sin v \\ z = \sin u \, ; \end{cases}$$

cette représentation paramétrique définit un difféomorphisme de $S_1 \times S_1$ sur le tore,

en désignant toujours par S_1 le cercle de rayon 1.

Plan tangent

Soit M un point d'une surface régulière (ou un point régulier d'une surface avec singularités). Si (φ, U) est une représentation paramétrique régulière au voisinage de M, l'image de $\mathbf{R}^2$ par l'application affine $T_M\varphi$ tangente à φ au point $m = \varphi^{-1}(M)$ (cf. chap. 1) est indépendante du choix de la représentation paramétrique régulière : cette image est le plan, passant par M, engendré par les vecteurs :

$$\frac{\partial \varphi}{\partial u}(m), \quad \frac{\partial \varphi}{\partial v}(m)$$

appelé plan tangent à la surface en M. On le désignera par $T_M S$. Ce plan tangent est l'ensemble engendré par les tangentes en M aux courbes régulières passant par ce point et tracées sur S. De manière précise, cela signifie que, pour tout arc paramétré régulier y tel que $\gamma(t_0) = M$ dont la trajectoire est contenue dans S, le vecteur vitesse $(d\gamma/dt)(t_0)$ d'origine M appartient au plan tangent; de plus, si $\varphi : U \rightarrow S$, $U \subset \mathbf{R}^2$ est une représentation paramétrique de S, l'application y se factorise localement sous la forme $y = \varphi \circ f$, où f est un arc paramétré de U, c'est-à-dire que l'arc paramétré est défini par $u = f_1(t)$, $v = f_2(t)$ si $f = (f_1, f_2)$. En particulier les vecteurs $(\partial\varphi/\partial u)(m)$ et $(\partial\varphi/\partial v)(m)$ correspondent respectivement à $v =$ constante et $u =$ constante. ***Réciproquement,*** pour tout vecteur $V \in T_M S$, il existe une application $f : I \rightarrow U$ telle que le vecteur vitesse :

$$\frac{d(\varphi \circ f)}{dt}(t_0)$$

soit égal au vecteur V.

Si la surface est réglée, toute génératrice passant par M appartient au plan tangent ; en particulier, s'il passe par M deux génératrices distinctes, elles engendrent le plan tangent : c'est ce qui se produit pour l'hyperboloïde à une nappe et pour le paraboloïde hyperbolique (fig. 10). Dans le cas particulier où le plan tangent est le même tout le long de chaque génératrice, on dit qu'on a une ***surface réglée développable.***

Examinons les différentes surfaces réglées développables. Si toutes les génératrices passent par un point fixe, on a un cône ; si elles sont parallèles à une direction ***fixe,*** on a un ***cylindre.*** Un autre exemple très important s'obtient à partir des tangentes à une courbe : l'ensemble des tangentes à une courbe régulière engendre une surface développable. On peut montrer que la surface développable la plus générale est formée de nappes de surfaces coniques, cylindriques et de tangentes à des courbes gauches, ces nappes étant attachées les unes aux autres le long d'une génératrice (deux nappes se « recollant » le long d'une génératrice ont même plan tangent).

Position par rapport au plan tangent

Au voisinage d'un point $M_0(x_0, y_0, z_0)$ régulier d'une surface S, on peut exprimer une des coordonnées en fonction des deux autres, par exemple $z = g(x, y)$. Le plan tangent en M_0 est alors défini par :

$$z - z_0 = \frac{\partial g}{\partial x}(m_0)(x - x_0) + \frac{\partial g}{\partial y}(m_0)(y - y_0),$$

où m_0 est le point (x_0, y_0) ; par un changement de repère euclidien, on peut ramener M_0 à l'origine des coordonnées, soit $x_0 = y_0 = z_0 = 0$, le plan tangent en ce point à S étant défini par $z = 0$, soit $(\partial g/\partial x)(m_0) = 0$ et $(\partial g/\partial y)(m_0) = 0$. Le nombre $z = g(x, y)$ représente alors la distance « algébrique » du point M $(x, y,$

$g(x, y))$ de la surface au plan tangent en M_0. La formule de Taylor donne :

$$z = g(x,y) = rx^2 + 2sxy + ty^2 + R_2,$$

où on a adopté les notations de Monge :

$$r = \frac{\partial^2 g}{\partial x^2}(m_0),\ s = \frac{\partial^2 g}{\partial x \partial y}(m_0),\ t = \frac{\partial^2 g}{\partial y^2}(m_0);$$

le reste R_2 est ici une quantité qui tend vers 0 plus vite que $x^2 + y^2$ quand (x, y) tend vers (0, 0).

Si la forme quadratique associée à la différentielle seconde de g :

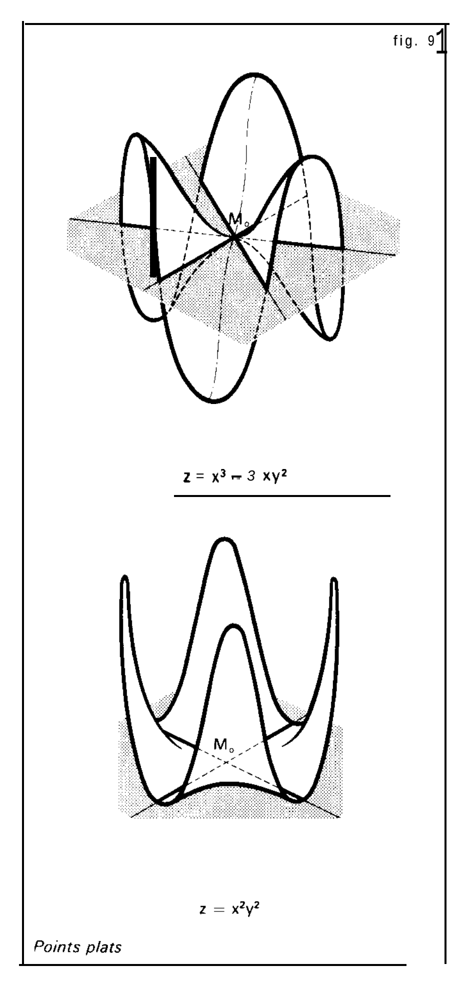
$$Q(x,y) = rx^2 + 2sxy + ty^2$$

est nulle ($r = s = t = 0$), on dit que le point M_0 est *plat* ; la surface et son plan tangent ont alors un contact d'ordre supérieur à 1. Par exemple, la surface d'équation $z = x^3 - 3xy^2$ (« selle de singe ») a un point plat à l'origine ; elle coupe son plan tangent à l'origine suivant trois droites et traverse ce plan tangent (fig. 9). Par contre, la surface $z = x^2y^2$, qui présente elle aussi un point plat à l'origine, coupe le plan tangent en ce point suivant deux droites et reste d'un même côté de ce plan tangent (fig. 9).

Si la forme quadratique Q est non nulle, on distingue trois cas (fig. 10) :

a) Si $rt - s^2 > 0$, la forme Q est définie positive ou négative. Au voisinage de M_0, l'intersection de S et du plan tangent en M_0 se réduit a M_0 et la surface reste d'un même côté de ce plan tangent (elle est localement convexe). C'est le cas de tous les points d'un ellipsoïde. On dit que le point M_0 est *elliptique*.

b) Si $rt - s^2 < 0$, la surface traverse son plan tangent au voisinage de M_0. C'est le cas de tout point d'un hyperboloïde à une nappe ou d'un paraboloïde hyperbolique. On dit que le point est *hyperbolique*.

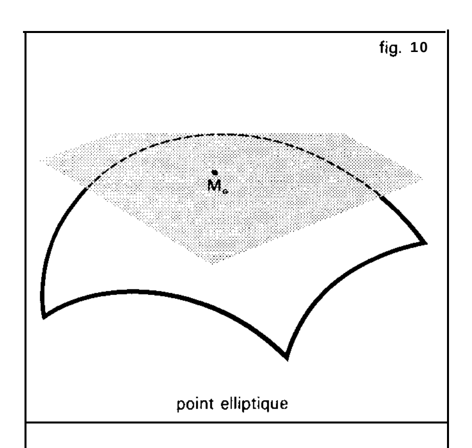


Points plats

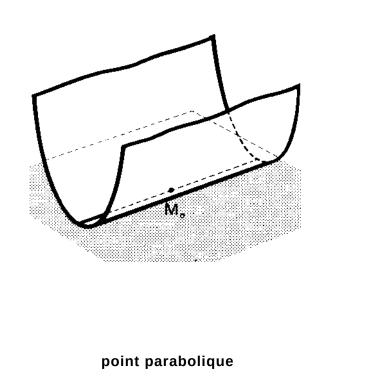
c) Si $rt - s^2 = 0$, on ne peut rien dire de général. Même si la surface reste localement d'un même côté du plan tangent (*point parabolique*), l'intersection ne se réduit pas nécessairement à un point. C'est le cas de tout point d'un cylindre elliptique ou parabolique ; l'intersection avec un plan tangent est ici une génératrice.

Une même surface peut présenter des points de nature différente. Par exemple, le tore, dont on a donné une représentation

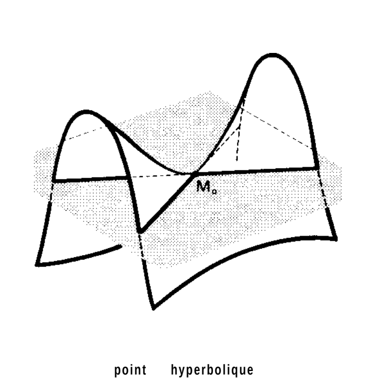
fig. 10



point elliptique



point parabolique

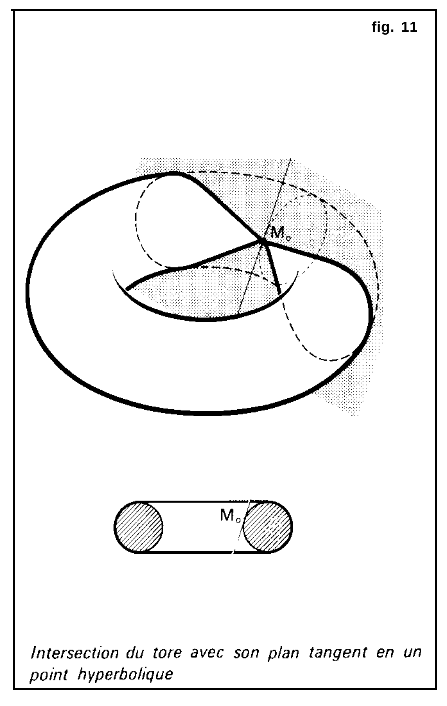


point hyperbolique

Position d'une surface par rapport à un plan tangent

paramétrique ci-dessus, admet des points elliptiques (tels que $x^2 + y^2 > 1$), des points hyperboliques (tels que $x^2 + y^2 < 1$; fig. 11) et des points paraboliques (tels que $x^2 + y^2 = 1$). L'ensemble des points paraboliques est constitué de deux parallèles au tore, le plus « haut » et le plus « bas », et en un tel point de plan tangent touche le tore tout le long du parallèle correspondant.

fig. 11



Intersection du tore avec son plan tangent en un point hyperbolique

6. Formes fondamentales sur une surface

On appelle *première forme fondamentale* sur une surface S la forme quadratique Φ qui, à tout vecteur V tangent à S en M. associe le carré de sa longueur, soit :

$$\Phi(V) = \|V\|^2 = V \cdot V, \quad V \in T_M S.$$

Si au voisinage de M la surface S admet pour représentation paramétrique $(u, v) \mapsto \varphi(u, v)$, on écrit :

$$D\varphi = \frac{\partial \varphi}{\partial u} du + \frac{\partial \varphi}{\partial v} dv,$$

et on a :

$$\Phi = D\varphi \,.\, D\varphi = E\, du^2 + 2\, F\, du\, dv + G\, dv^2,$$

en posant :

$$E = \left(\frac{\partial \varphi}{\partial u}\right), \quad F = \frac{\partial \varphi}{\partial u} \cdot \frac{\partial \varphi}{\partial v}, \quad G = \left(\frac{\partial \varphi}{\partial v}\right)^2;$$

justifions ces notations. Si y est un arc paramétré, dont le vecteur vitesse en M est égal à V, qui se factorise sous la forme $y = \varphi \circ f$ (avec les notations du chapitre précédent), alors on a :

$$V = \left(\frac{\partial \varphi}{\partial u}\frac{du}{dt} + \frac{\partial \varphi}{\partial v}\frac{dv}{dt}\right)_{t_0} = D\varphi\left(\frac{d\gamma}{dt}(t_0)\right),$$

et par suite :

$$\Phi(V) = D\varphi\left(\frac{d\gamma}{dt}(t_0)\right) \cdot D\varphi\left(\frac{d\gamma}{dt}(t_0)\right).$$

Définissons maintenant la deuxième forme fondamentale ; il sera pour cela nécessaire d'orienter la surface. Si la représentation paramétrique $\varphi : U \longrightarrow S$, $U \subset \mathbf{R}^2$ est régulière en tout point M de $\varphi(U)$, le produit vectoriel :

$$\frac{\partial \varphi}{\partial u}(m) \wedge \frac{\partial \varphi}{\partial v}(m),$$

où $m = \varphi^{-1}(M) \in U$, est non nul ; on peut donc associer à chaque point M de $\varphi(U)$ un vecteur unitaire n normal à S de même sens que le produit vectoriel précédent, c'est-à-dire on *oriente* la surface. Si un arc paramétré :

$$\gamma : I \rightarrow S, I \subset \mathbf{R},$$

se factorise sous la forme $y = \varphi \circ f$ (cf. *infra*), alors le produit scalaire :

$$n \cdot \frac{d^2\gamma}{dt^2}(t_0) = n \cdot D^2\varphi\left(\frac{d\gamma}{dt}(t_0), \frac{d\gamma}{dt}(t_0)\right),$$

où $D^2\varphi$ est la dérivée seconde de φ, ne dépend que du vecteur tangent :

$$V = \frac{d\gamma}{dt}(t_0) = D\left(\frac{df}{dt}(t_0)\right);$$

en effet, on a :

$$\frac{d^2\gamma}{dt} = \frac{\partial \varphi}{\partial u}\frac{d^2u}{dt^2} + \frac{\partial \varphi}{\partial v}\frac{d^2v}{dt^2} + \frac{\partial^2 \varphi}{\partial u^2}\left(\frac{du}{dt}\right)^2 + 2\frac{\partial^2 \varphi}{\partial u \partial v}\frac{du}{dt}\frac{dv}{dt} + \frac{\partial^2 \varphi}{\partial^2 v}\left(\frac{dv}{dt}\right)^2$$

d'où :

$$\psi(V) = n \cdot D^2\varphi\left(\frac{d\gamma}{dt}(t_0), \frac{d\gamma}{dt}(t_0)\right) = \mathcal{L}\left(\frac{du}{dt}\right)^2 + 2\,\mathcal{M}\left(\frac{du}{dt}\right)\left(\frac{dv}{dt}\right) + \mathcal{N}\left(\frac{dv}{dt}\right)^2,$$

en posant :

$$\mathcal{L} = n \cdot \frac{\partial^2 \varphi}{\partial u^2}, \quad \mathcal{M} = n \cdot \frac{\partial^2 \varphi}{\partial u \partial v}, \quad \mathcal{N} = n \cdot \frac{\partial^2 \varphi}{\partial v^2}.$$

La forme quadratique $\psi : V \mapsto \psi(V)$ est appelée la deuxième forme fondamentale de la surface. On démontre que la quantité :

$$K = \frac{\mathcal{L}\mathcal{N} - \mathcal{M}^2}{EG - F^2}$$

est indépendante de la représentation paramétrique. En particulier, pour la représentation paramétrique $z = g(x, y)$ utilisée dans le chapitre précédent (avec $p = q = 0$ en M), on a :

$$K = rt - s^2;$$

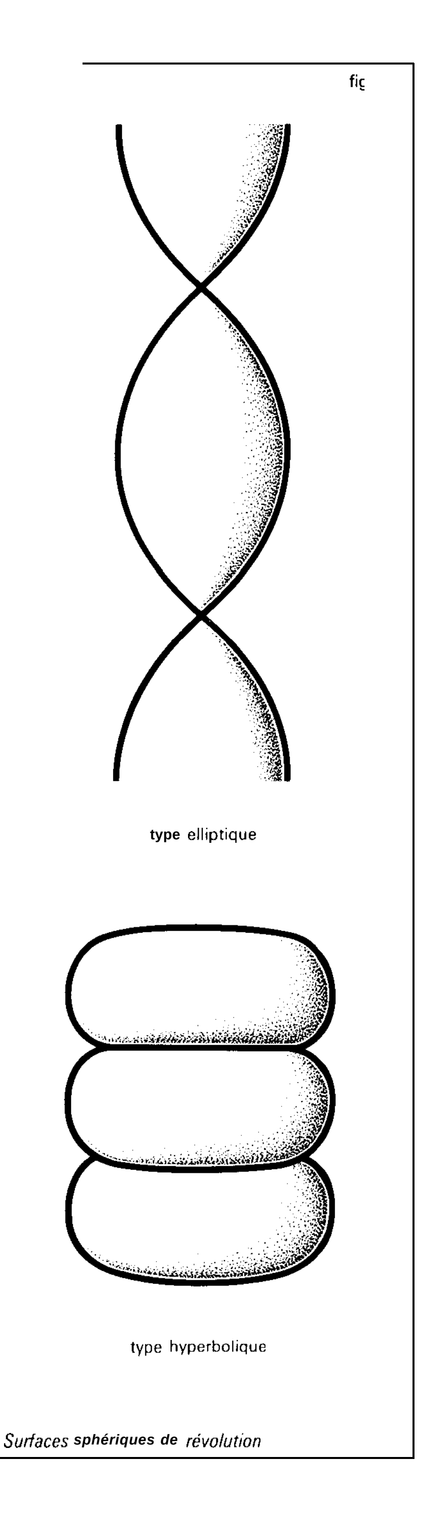
ce scalaire K s'appelle la *courbure totale,* ou courbure de Gauss, en M. Suivant le signe de K, le point est elliptique, hyperbolique ou parabolique (cf. chap. 5, Posi-

tion par rapport au plan tangent) ; remarquons que, puisque la forme quadratique Φ est définie positive, le signe de K est celui de Ch" $\mathcal{M}^2$.

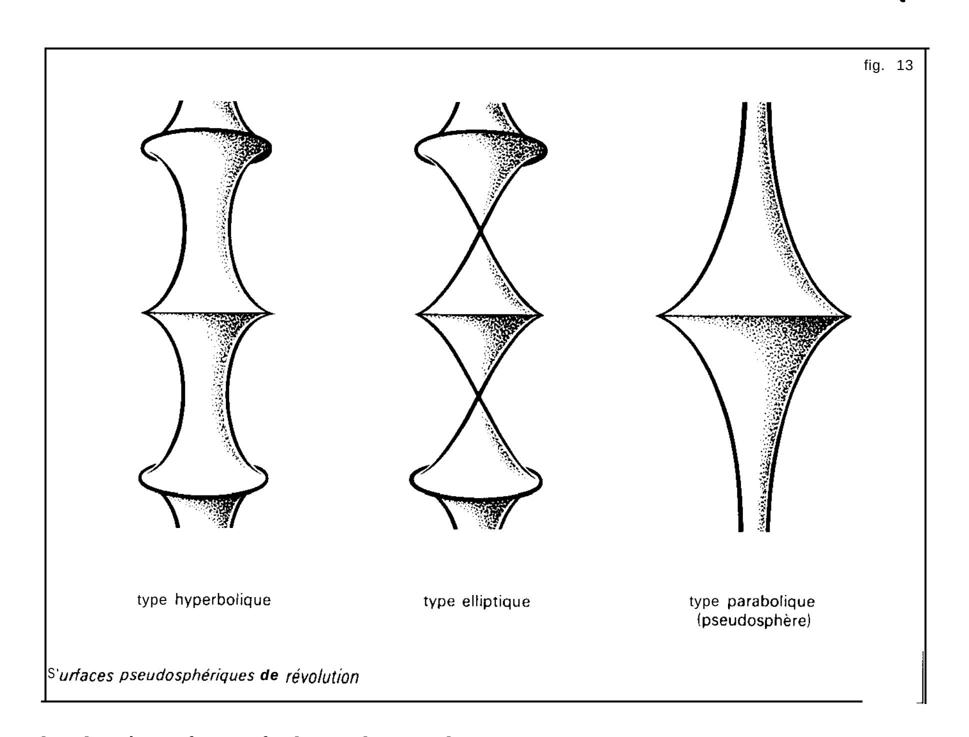
C. Gauss a démontré que la courbure totale était déterminée par E, F, G et leurs dérivées partielles premières. Étant donné deux surfaces S et S', on appelle ***isométrie locale*** de S dans S' un difféomorphisme d'un ouvert U de $\mathbf{E}_3$ dans $\mathbf{E}_3$, appliquant $S \cap U$ dans S' et transformant en chaque point M de $S \cap U$ la première forme fondamentale de S en la première forme fondamentale de S' ; par suite, une isométrie locale laisse invariante la courbure totale. En particulier, une surface S telle qu'il existe en chaque point $M \in S$ une isométrie d'un voisinage de ce point dans S sur un ouvert plan est dite ***applicable sur le plan*** ; puisque, pour un plan, la deuxième forme fondamentale est nulle, la courbure totale d'une surface applicable sur un plan est nulle. On démontre que toute surface à courbure totale nulle est une surface développable et que toute surface développable est applicable sur le plan.

Plus généralement, on montre que si deux surfaces S et S' ont une courbure totale constante (cette constante étant la même pour les deux surfaces). il existe des isométries locales de S sur S'. Si la courbure totale est une constante positive, on dit que la surface est une surface sphérique ; parmi ces surfaces figurent la sphère, et les surfaces sphériques de révolution du type hyperbolique et du type elliptique (fig. 12). Si la courbure totale est constante négative. on a les surfaces pseudosphériques ; celles qui sont de révolution se répartissent en trois types indiqués par la figure 13.

Enfin, signalons qu'on démontre qu'une isométrie locale qui conserve aussi



Surfaces **sphériques de** *révolution*



Surfaces pseudosphériques de révolution

la deuxième forme fondamentale est la restriction d'un déplacement euclidien ; par suite, l'ensemble des deux formes fondamentales caractérise localement une surface.

7. Courbes tracées sur une surface

Soit C une courbe régulière orientée tracée sur une surface régulière S ; à tout point M de C on va attacher un repère, appelé *trièdre de Darboux*, obtenu de la manière suivante : soit ***t***, ***n***, ***b*** le trièdre de Frénet de la courbe C au point M ; le trièdre de Darboux e_1, e_2, e_3 s'obtient en prenant pour e_3 le vecteur unitaire normal en M à la surface associé à l'orientation de cette surface, et en prenant $e_1 = t$ (et $e_2 = e_3 \wedge e_1$ pour obtenir un trièdre direct). Si s est l'abscisse curviligne sur C, on a alors les formules :

$$\frac{de_1}{ds} = \frac{1}{\rho_g} e_2 + \frac{1}{\rho_n} e_3,$$

$$\frac{de_2}{ds} = -\frac{1}{\rho_g} e_2 + \frac{1}{\tau_g} e_1,$$

$$\frac{de_3}{ds} = -\frac{1}{\rho_n} e_2 + \frac{1}{\tau_g} e_3,$$

les nombres $1/\rho_g$, $1/\rho_n$ et $1/\tau_g$ ainsi définis s'appellent respectivement la *courbure géodésique*, la *courbure normale* et la *torsion géodésique* en M. Si ρ et τ sont la courbure et la torsion de C en M, on a, en désignant par θ l'angle des deux vecteurs ***n*** et e_3 :

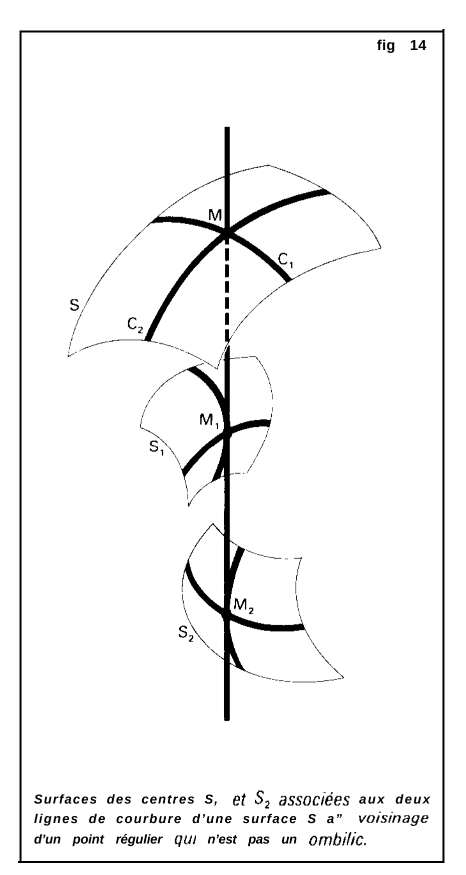
$$\frac{1}{\rho_g} = \frac{\sin\theta}{\rho}, \quad \frac{1}{\rho_n} = \frac{\cos\theta}{\rho}, \quad \frac{1}{\tau_g} = \frac{d\theta}{ds} + \frac{1}{\tau},$$

La courbure normale $1/\rho_n$ en M est la même pour toutes les courbes tracées sur

S qui admettent la même tangente en ce point M ; en effet, $1/\rho_n = e_3.\ (d^2M/ds^2)$, ce qui entraîne que cette courbure normale est égale à $\psi(e_1)$, en désignant par ψ la deuxième forme fondamentale. On en déduit *le théorème de Meusnier* : Si un plan P pivote autour d'une droite du plan tangent T_MS, alors le centre de courbure (en M) de la section de S par P décrit un cercle passant par le point M. Un point d'une surface S pour lequel la courbure normale est la même dans toutes les directions est appelé un *ombilic.*

Une courbe tracée sur S est appelée une *ligne asymptotique* si la courbure normale en chacun de ses points est nulle ; en chaque point, tout vecteur tangent à une telle courbe annule la deuxième forme fondamentale. Par suite, il ne passe de lignes asymptotiques que par les points hyperboliques (deux directions asymptotiques) ou paraboliques (une direction asymptotique). Toute droite tracée sur une surface est une ligne asymptotique.

On appelle *lignes de courbure* les lignes dont la torsion géodésique en chaque point est nulle, et on montre que cette condition équivaut à dire que les normales à la surface S le long d'une telle courbe engendrent une surface développable ; ces normales sont les tangentes à la courbe décrite par les centres de courbure de la ligne de courbure. En chaque point régulier qui n'est pas un ombilic, il existe deux directions du plan tangent, appelées *directions principales,* perpendiculaires entre elles, pour lesquelles la courbure normale atteint son maximum $1/R_1$ et son minimum $1/R_2$; les lignes de courbure peuvent aussi être définies commc les courbes tangentes en chaque point aux directions principales (fig. 14). La courbure totale est reliée



Surfaces des centres S_1 et S_2 associées aux deux lignes de courbure d'une surface S a" voisinage d'un point régulier qui n'est pas un ombilic.

simplement aux courbures des lignes de courbure par la formule :

$$K = \frac{1}{R_1 R_2}.$$

Sur une surface dont tous les points sont des ombilics, toutes les courbes sont lignes de courbure ; les seules surfaces possédant cette propriété sont les plans (K = 0) et les sphères (K = $1/R^2$ constant).

On appelle *géodésiques* les courbes dont la courbure géodésique est nulle, c'est-à-dire dont la normale principale est normale à la surface ; les géodésiques sont définies par un système d'équations différentielles du second ordre, et on démontre

que par tout point d'une surface il passe une géodésique et une seule admettant une tangente donnée. De même, par deux points M et M' de S assez voisins passe une géodésique et une seule : la longueur de l'arc de géodésique MM' réalise alors le minimum de la longueur des arcs joignant M à M'. Par exemple, les géodésiques du plan sont les droites ; les géodésiques d'une sphère sont les grands cercles, sections de la sphère par les plans passant par le centre, et par deux points non diamétralement opposés passe un tel grand cercle et un seul. Les géodésiques d'un cylindre de révolution sont les parallèles et les hélices circulaires.

On montre que la courbure géodésique ne dépend que de la première forme fondamentale ; par suite, une isométrie locale applique les géodésiques sur les géodésiques.

8. Propriétés globales liées à la courbure totale

Soit y : I → S un arc paramétré d'une surface S. Si X = X(t) est un champ de vecteurs le long de la courbe C = y(I), on définit la *dérivée* ***covariante*** DX/dt du champ X au point M = $\gamma(t)$ en projetant le vecteur dX/dt sur le plan tangent T_MS parallèlement à la normale. On dit alors que le champ X ***se*** *déplace* ***par parallélisme,*** ou est parallèle, si pour tout $t \in 1$ la dérivée covariante est nulle. Remarquons que la valeur $X(t_0)$ du champ en un point détermine alors le champ parallèle. En particulier, on dit qu'un arc paramétré est géo***désique*** si sa vitesse se déplace par parallélisme ; une courbe géodésique devient un arc géodésique si on prend pour paramètre l'abscisse curviligne s ou tout autre paramètre t = ***as*** + b, ***a*** et b constants avec ***a*** $\neq$ ***0.***

Si la courbure totale n'est pas nulle, le transport par parallélisme le long d'un ***lacet*** (c'est-à-dire un arc paramétré y : ***[a,*** b] → S tel que ***y(a)*** = y(b) ne ramène pas en général le vecteur $X(a)$ à sa position initiale et, par suite, si on considère deux points M et M', le transport par parallélisme de M à M' dépend du chemin choisi. En effet, on démontre que la « variation de l'angle » du vecteur X par transport parallèle le long d'un lacet est l'intégrale :

$$\iint_\Sigma \frac{d\sigma}{R_1 R_2},$$

où Σ est la partie de S limitée par le lacet et $d\sigma$ l'élément d'aire sur S.

D'autre part, on montre que si le lacet y se compose d'un nombre fini d'arcs différentiables γ_k séparés par des points anguleux où l'angle du vecteur tangent à y subit une discontinuité 8,. on a :

$$\sum_i \int_{\gamma_i} \frac{ds}{\rho_g} + \sum_i \theta_i = 2\pi - \iint_\Sigma \frac{d\sigma}{R_1 R_2},$$

formule de Gauss-Bonnet. Dans le cas particulier ***où y*** est un ***triangle géodésique,*** c'est-à-dire un triangle curviligne dont les côtés sont des arcs géodésiques, l'intégrale de la torsion géodésique est nulle. Si on désigne par α_1, α_2, α_3 les mesures en radian (comprises entre 0 et 2π) des angles du triangle, on a, puisque ces angles sont les supplémentaires de ceux qui interviennent dans la formule de Gauss-Bonnet :

$$\alpha_1 + \alpha_2 + \alpha_3 = \pi + \iint \frac{d\sigma}{R_1 R_2};$$

pour le plan on retrouve le résultat classique que la somme des angles d'un triangle

est égale à π. Sur une sphère de rayon R, on a :

$$\alpha_1 + \alpha_2 + \alpha_3 = \pi + \frac{A}{R_2},$$

où A est l'aire du triangle. Remarquons que si la courbure totale est négative en tout point de S, alors la somme des angles d'un triangle géodésique est inférieure à π. On retrouve que les surfaces à courbure constante constituent des modèles pour les géométries non euclidiennes de Riemann et de Lobatchewski.

Si on considère maintenant une surface compacte (c'est-a-dire fermée et bornée) sans bord, on montre qu'on peut la trianguler, c'est-à-dire la découper en domaines limités par des triangles curvilignes (pas nécessairement géodésiques). Appliquant la formule de Gauss-Bonnet à chaque triangle et faisant la somme, on obtient, puisque chaque arc est parcouru deux fois en sens contraires :

$$\iint_S \frac{d\sigma}{R_1 R_2} = 2\pi(n_2 - n_1 + n_0),$$

où n_2, n_1 et n_0 sont respectivement le nombre de triangles, le nombre d'arêtes et le nombre de sommets de la triangulation. Le premier membre étant indépendant de la triangulation, il en est de même du second. Le nombre entier positif :

$$n_2 - n_1 + n_0$$

est appelé la *caractéristique d'Euler-Poincaré* de la surface ; par exemple, pour la sphère il est égal à 2, car :

$$\iint_S \frac{d\sigma}{R^2} = 4\pi;$$

pour le tore, la caractéristique d'Euler-Poincaré est nulle.

On montre que la caractéristique d'Euler-Poincaré est un invariant topologique de la surface. On montre aussi qu'il existe, sur une surface S, un champ différentiable de vecteurs tangents ne s'annulant en aucun point si, et seulement si, la caractéristique d'Euler-Poincaré est nulle ; il n'existe donc pas de tel champ sur une sphère.

PAULETTE LIBERMANN

Bibliographie

J.-M. BRAEMER & Y. KERBRAT, *Géométrie des courbes et des surfaces*, Hermann, Paris, 1976 / H. CARTAN, *Formes différentielles*, *ibid.*, 1967 / P. DOMBROWSKI, *150 Years after Gauss' « Disquisitiones genereales circa superficies curvas »*, astérisque 62, Société mathématique de France, Paris, 1979 / L. P. EISENHART, *An Introduction to Differential Geometry with Use of the Tensor Calculus*, repr. of 1947, Books on Demand, Ann Arbor (Mich.) / D. LEBORGNE, *Calcul différentiel et géométrie*, P.U.F., 1982 / P. LIBERMANN, « Géométrie différentielle », in J. Dieudonné et al., *Abrégé d'histoire des mathématiques*, t. II, Hermann, 1978 / J. PICHON, *Les Courbes dans le plan et dans l'espace*, Ellipses, 1987 / 1. PORTEOUS, *Geometric Differentiation*, Cambridge Univ. Press, New York, 1992 / M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, 5 vol., Publish or Perish, Houston (Texas), 1979 / S. STERNBERG, *Lectures on Differential Geometry*, Chelsea Publ., New York, 2' éd. 1983 / J. J. STOKER, *Differential Geometry*, Wiley-Interscience, New York, 1989 / P. THUILLIER, J.-C. BELLOC & A. DE VILLÈLE, *Mathématiques géométrie différentielle*, Masson, Paris, 2' éd. 1991.

GROUPES

LES IDÉES de *symétrie* et de *régularité se* retrouvent dans toutes les civilisations, bien avant que ne fût conçue la notion de groupe : par exemple, presque tous les groupes discrets de déplacements du plan (il y en a dix-sept types non isomorphes) sont sous-jacents aux multiples ornements géométriques imaginés par les artistes

arabes. Les Grecs, dans leur géométrie, ont été très tôt intéressés par les propriétés de régularité, et on sait que le couronnement des *Éléments* d'Euclide est la construction des cinq polyèdres réguliers, ce qui, en substance, revient à la détermination des groupes finis de rotations dans l'espace à trois dimensions.

Toutefois, la notion de groupe n'apparaît explicitement qu'au cours des travaux sur la résolution des équations algébriques « par radicaux », au début du XIX[e] siècle ; développant une idée de Lagrange, Ruffini et Cauchy sont amenés à considérer les groupes de permutations des racines d'une équation algébrique qui laissent invariantes certaines fonctions de ces racines ; et c'est en approfondissant cette idée que Galois obtiendra ses résultats décisifs sur la résolution par radicaux. Ces premiers groupes sont donc des groupes *finis*, et c'est sous la forme de la théorie des groupes de permutations que la théorie générale des groupes finis commencera à se développer (notamment chez Mathieu et Jordan) jusque vers 1870. Les débuts de la cristallographie mathématique (vers 1830) font apparaître d'autres groupes finis, cette fois formés de rotations et de symétries laissant un point fixe ; enfin, Jordan, en 1868, aborde franchement l'étude des groupes de déplacements (finis ou non) dans l'espace euclidien à trois dimensions. Un peu plus tard, Klein et Poincaré feront des groupes de déplacements non euclidiens le fondement de leur théorie des fonctions automorphes, tandis que Lie, cherchant à réaliser pour les équations différentielles ce que Galois avait fait pour les équations algébriques, crée la théorie générale des groupes continus de transformations (actuellement appelés groupes de Lie). En même temps, Klein est amené, par ses réflexions sur les fondements de la géométrie « élémentaire », à mettre la notion de groupe de transformations à la base même de cette branche des mathématiques, qui devient un simple chapitre de la théorie des groupes classiques développée depuis Jordan pendant toute la fin du XIX[e] siècle.

Il faut attendre la fin du XIX[e] siècle pour que la structure de groupe telle que nous la concevons aujourd'hui soit enfin définie de façon intrinsèque (et non plus en se restreignant au cas où les éléments du groupe sont des transformations). Depuis lors, la notion de groupe a envahi toutes les mathématiques contemporaines. On s'est, d'une part, aperçu du caractère protéiforme de l'idée de groupe, débordant largement le concept initial de groupe « ensembliste » (groupes topologiques, groupes algébriques, schémas en groupes et, plus généralement, « objets en groupes » d'une catégorie représentant un foncteur représentable de cette catégorie dans la catégorie des groupes) ; on a, en outre, découvert de surprenantes relations entre des types de groupes très divers (par exemple entre les groupes de Lie, les groupes algébriques, les groupes « arithmétiques » et les groupes finis). D'autre part, l'expérience a montré l'extraordinaire efficacité de la notion de groupe dans toutes les parties des mathématiques, une fois qu'on parvient à l'y introduire : groupes d'homologie et d'homotopie en topologie algébrique, espaces fibrés principaux en géométrie différentielle et en topologie différentielle en sont des exemples bien connus ; un autre exemple, plus remarquable encore, est la possibilité de définir une structure de groupe sur l'ensemble des classes de structures différentielles compatibles avec une variété (topologique) donnée. Cette tendance a gagné la physique elle-même : en cherchant

à expliquer les symétries expérimentales qu'ils constataient dans les phénomènes atomiques, les théoriciens se sont naturellement tournés vers la théorie des groupes, avec un succès assez remarquable, bien que fort mystérieux.

JEAN DIEUDONNÉ



A. Généralités

On se propose de présenter ici les notions fondamentales de théorie des groupes qui interviendront constamment dans la suite. Celle-ci contient un très grand nombre d'exemples, c'est pourquoi cet exposé introductif n'explicite que quelques groupes utilisés aussi ailleurs, notamment en cristallographie, en chimie, en linguistique.

1. La structure de groupe

Un groupe G est un ensemble muni d'une loi de composition interne :

$$(x, y) \mapsto x * y$$

qui possède les propriétés suivantes :

(a) Elle est *associative*, c'est-à-dire que, si a, b, c sont des éléments de G, on a :

$$a * (b * c) = (a * b) * c\,;$$

(b) Elle admet un *élément neutre*, c'est-à-dire qu'il existe un élément $e \in$ G (nécessairement unique, manifestement) tel que, pour tout a $\in$ G :

$$a * e = e * a = a\,;$$

(c) Tout élément a de G admet un *symétrique* (en notation multiplicative on dira un inverse), c'est-à-dire qu'il existe un élément de G, noté a^{-1}, tel que :

$$a^{-1} * a = a * a^{-1} = e.$$

On ne se préoccupera pas ici de savoir si l'on peut affaiblir ces axiomes en jonglant avec des hypothèses « à droite » et « à gauche » dans (b) et dans (c). Le groupe est dit *commutatif*, *ou abélien,* si la loi de composition est commutative, c'est-à-dire $a * b = b * a$ pour tout couple d'éléments de G. Cette loi est alors souvent (mais pas toujours) notée additivement, par le signe + ; l'élément neutre est désigné par 0 et le symétrique d'un élément a est noté $-a$. C'est le cas, par exemple, pour la loi de groupe sous-jacente à une structure d'anneau ou d'espace vectoriel. On appelle *ordre* d'un groupe fini G le nombre G de ses éléments.

Dans ce qui suit, sauf mention explicite d'une autre notation, la notation multiplicative sera adoptée systématiquement, ce qui signifie que l'on notera $x.$ y, ou plus simplement xy, l'image du couple (x, y) par la loi de composition. L'élément neutre sera désigné par 1. Lorsque plusieurs groupes seront considérés simultanément, le même symbole 1 désignera donc *plusieurs* objets mathématiques distincts, ce qui paraît en contradiction avec les règles logiques les plus simples (cf. par exemple la formule (2) ci-dessous) ; en fait, cela n'est guère gênant, car le contexte mathématique permet toujours d'éviter toute ambiguïté. Ainsi, dans la formule (2), puisquefest une application de G dans G', le symbole 1 dans la partie gauche de la formule représente l'élément neutre de G tandis que le 1 de droite représente l'élément neutre de G'.

Remarquons maintenant que, si l'on multiplie à gauche par a^{-1} les deux membres de l'égalité $ax = ay$, on obtient, en

applicant l'associativité, $1x = 1y$, d'où $x = y$. On a ainsi obtenu la règle ***de simplification*** dans un groupe : si ***a, x, y*** sont des éléments d'un groupe, on a les équivalences :

$$ax = ay \Leftrightarrow xa = ya \Leftrightarrow x = y.$$

Une démonstration tout à fait analogue montre que, dans un groupe, ***les équations linéaires,*** du type $ax = b$, ***ou*** $xa = b$, ont toujours une solution unique ; par multiplication à gauche par a^{-1}, on obtient par exemple que la première a pour solution $x = a^{-1}b$.

Si x est un élément d'un groupe G et n un entier positif, on notera x^n le produit de n éléments égaux à x, et x^{-n} le produit de n éléments égaux à x^{-1}. L'élément x^0 étant par définition l'élément neutre, on a donc défini x^n pour tout entier relatif n et on vérifie facilement que deux puissances quelconques d'un même élément commutent toujours et que :

$$x^n x^m = x^{m+n}, \quad m, n \in \mathbf{Z} ;$$

en notation additive on écrit nx au lieu de x^n.

Morphismes

Conformément **aux** définitions générales pour les structures algébriques, on dit qu'une application f d'un groupe G dans un groupe G' est un ***morphisme, ou*** un ***homomorphisme,*** de groupe si on a :

$$(1) \qquad f(xy) = f(x)f(y)$$

pour tout couple d'éléments de G. Par exemple, le logarithme usuel réalise un homomorphisme du groupe multiplicatif $\mathbf{R}_+^*$ des nombres réels strictement positifs sur le groupe additif de tous les nombres réels, car :

$$\ln xy = \ln x + \ln y, \; x, y \in \mathbf{R}_+^* ;$$

bien entendu, il faut, quand les deux groupes ne sont pas tous les deux notés multiplicativement, adapter les notations de la condition (1). Un morphisme bijectif est appelé un ***isomorphisme*** ; c'est le cas du logarithme qui réalise un isomorphisme du groupe multiplicatif RT sur le groupe additif R. Dans le cadre de la théorie des groupes, il n'y a pas lieu de distinguer des groupes isomorphes, et on parlera parfois (par abus de langage) de réalisation d'un même groupe pour désigner des groupes isomorphes. Remarquons enfin que, si on prend pour y l'élément neutre de G dans (1), on obtient, après simplification :

$$(2) \qquad f(1) = 1,$$

qui montre que tout morphisme de G dans G' transforme l'élément neutre de G en l'élément neutre de G'. Les groupes, et leurs morphismes, forment un exemple très simple de catégorie.

Sous-groupes

Une partie non vide H d'un groupe G est un ***sous-groupe*** si le composé de deux éléments de H est encore un élément de H et si H est un groupe pour la loi de composition ainsi définie ; on vérifie facilement qu'une partie H non vide d'un groupe G est un sous-groupe si et seulement si $xy^{-1} \in$ H pour tout couple (x, y) d'éléments de H. Des exemples très simples de sous-groupes s'obtiennent à partir des morphismes : si f : G $\rightarrow$ G' est un morphisme de groupe, alors son *image* $f(\mathrm{G})$ est un sous-groupe de G' et son ***noyau*** Ker $f = f^{-1}(1)$ est un sous-groupe de G (en fait, comme on le verra ci-dessous au chapitre 3, le noyau n'est pas n'importe quel sous-groupe). Si f : G $\rightarrow$ G' et $g : \mathrm{G}' \rightarrow \mathrm{G}''$ sont deux morphismes, on dira que la « suite » :

$$\mathrm{G} \xrightarrow{f} \mathrm{G}' \xrightarrow{g} \mathrm{G}''$$

est *exacte* si l'image defest égale au noyau de g ; cette situation est fondamentale en algèbre homologique.

Il est clair que l'intersection d'une famille quelconque de sous-groupes est encore un sous-groupe (éventuellement le sous-groupe { 1 } réduit à l'élément neutre). Par suite, si K est une partie quelconque d'un groupe G, il existe un « plus petit » sous-groupe contenant K, à savoir l'intersection de tous les sous-groupes contenant K ; si H est ce sous-groupe, on dit qu'il est *engendré* par K, ou encore que K est un *système de générateurs* de H. Les éléments de H sont les produits finis $x_1x_2 \ x_n$, où l'un au moins des deux éléments x_i ou x_i^{-1} appartient à K ; en effet, tout sous-groupe contenant K contient ces éléments et l'ensemble de ces éléments est un groupe.

Indiquons enfin que, si A et B sont deux parties d'un groupe G, on note AB l'ensemble des produits ah pour a $\in$ A et $b \in$ B.

2. Quelques exemples

Dans de nombreux cas, les éléments d'un groupe G seront réalisés comme des bijections d'un ensemble E sur lui-même ; par définition, le produit ab de deux telles bijections est alors la bijection composée obtenue en faisant d'abord b, puis a. L'ensemble $\Sigma(E)$ de toutes les bijections d'un ensemble E est un groupe, appelé le *groupe symétrique* de l'ensemble E, pour la loi de composition ainsi définie.

Groupes cycliques

Un groupe G est dit *cyclique* s'il est engendré par un de ses éléments a. Tout élément de G est ainsi une puissance de a, et G est donc commutatif. Par exemple, le groupe additif Z des entiers relatifs est engendré par l'élément 1 ; car, avec les notations ci-dessus, n = nl. Si G est un groupe cyclique quelconque (on revient à la notation multiplicative), engendré par $a \in$ G, l'application :

$$n \mapsto a^n$$

est un morphisme surjectif de Z sur G. Si ce morphisme est injectif, c'est un isomorphisme. Dans le cas contraire, il existe des entiers n et n' distincts tels que $a^n = a^{n'}$; si on suppose $n' > n$, on en déduit $a^{n-n'} = 1$. Désignons par p le plus petit entier positif tel que $a^p = 1$; les éléments :

$$a^0 = 1, a, a^2, a^{p-1}$$

sont donc distincts et ce sont les seuls éléments du groupe G, car pour tout entier n on a : $a^n = a^{pq+r} = a^r$ si $n = pq + r$ est l'identité de division euclidienne de n par p, avec $r \in \{0, 1, ..., p-1\}$. Ainsi tout groupe cyclique infini est isomorphe à Z ; tous les groupes cycliques finis de même ordre p sont isomorphes entre eux. On désignera le groupe cyclique d'ordre p par C_p. On peut le réaliser comme l'ensemble des rotations du plan de centre 0 et d'« angles » $2k\pi/p$, $k = 0, 1, ... p$ 1. la loi de groupe étant la composition des rotations, ou encore comme l'ensemble des rotations d'angle $2k\pi/p$ autour de l'axe Oz dans l'espace à trois dimensions. Remarquons que le groupe multiplicatif des racinesp-ièmes de l'unité dans le corps des nombres complexes (cf. nombres COMPLEXES) est aussi une réalisation de ce groupe.

Groupes diédraux

Pour $n \geqslant 3$, on appelle *groupe diédral* D_n le groupe des rotations et des symétries du plan qui conservent un polygone régulier à n sommets. Ce groupe est d'ordre 2 n, car il contient n rotations, qui forment un

sous-groupe isomorphe au groupe cyclique C_n et n symétries (par rapport aux n droites joignant les sommets au centre du polygone). Si on numérote les sommets 1, 2, ... n (en choisissant un « sens de parcours » sur le polygone), le groupe D_n est engendré par la rotation a :

$$\begin{matrix} 1 & 2 & 3 & \dots & n-1 & n \\ \downarrow & \downarrow & \downarrow & & \downarrow & \downarrow \\ 2 & 3 & 4 & & n & 1 \end{matrix}$$

et la symétrie b :

$$\begin{matrix} 1 & 2 & 3 & \dots & n-1 & n \\ \downarrow & \downarrow & \downarrow & & \downarrow & \downarrow \\ 1 & n & n-1 & \dots & 3 & 2 \end{matrix}$$

autour de la droite joignant 1 au centre du polygone. Les générateurs a et b vérifient les « relations » :

(3) $$a^n = 1,\ b^2 = 1,\ ab = ba^{-1};$$

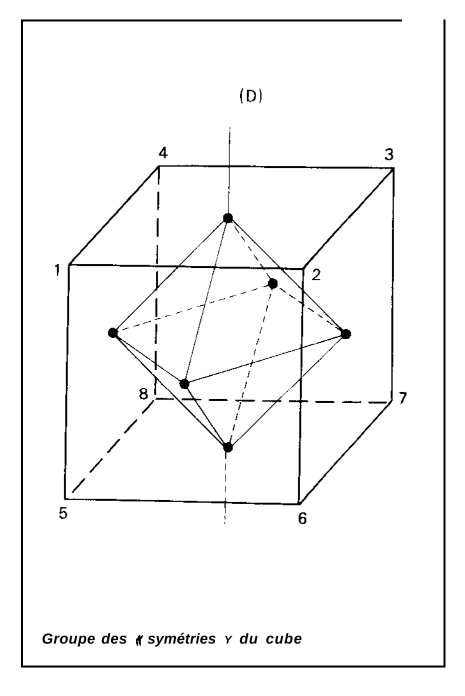
il en résulte que tout élément du groupe est de la forme a^k si c'est une rotation, ou de la forme $a^k b$ si c'est une symétrie, avec $k = 0, 1, ..., n-1$. Ces relations déterminent entièrement le groupe D,,.

On peut donner des réalisations de D_n comme groupe de déplacements de l'espace à trois dimensions, par exemple en prenant pour rotations des rotations autour de l'axe Oz et pour symétries des symétries autour de n droites du plan xOy faisant entre elles des angles égaux ; on peut obtenir une autre réalisation en remplaçant les symétries précédentes par des symétries autour de n plans passant par 0:. En cristallographie, on considère aussi le groupe D_{nh} d'ordre $4n$ des déplacements de l'espace à trois dimensions qui conservent un polygone régulier à n sommets du plan xOy ; on peut le réaliser comme le groupe engendré par D_n (dans la réalisation précédente) et la symétrie par rapport à l'origine.

Pour $n = 2$, les relations (3) définissent un groupe commutatif d'ordre 4, dont les éléments sont 1, a, b, ab, avec $a^2 = b^2 = 1$, ab = bu. C'est le 4-groupe de Klein. On peut le réaliser comme le groupe des symétries qui conservent un rectangle (qui n'est pas un carré) ; les nombreuses réalisations intuitives que l'on peut donner de ce groupe lui donnent une grande importance dans la pédagogie et l'enseignement élémentaire des mathématiques.

Les « symétries » du cube

Considérons un cube dont les sommets sont numérotés comme l'indique la figure.



Groupe des « symétries » du cube

Le groupe G des déplacements conservant le cube est engendré par la rotation a :

$$\begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 4 & 1 & 6 & 7 & 8 & 5 \end{matrix}$$

autour de la droite D joignant les centres des deux faces opposées l-2-3-4 et 5-6-7-8, par la rotation b :

1	2	3	4	5	6	1	8
↧	↧	↧	↧	↧	↧	↧	↧
1	4	8	5	2	3	7	6

autour de la diagonale l-7 et par la symétrie c :

1	2	3	4	5	6	7	8
↧	↧	↧	↧	↧	↧	↧	↧
5	6	7	8	1	2	3	4

autour du plan équidistant des deux faces opposées ci-dessus. Ce groupe est d'ordre 48. Les rotations forment un sous-groupe G, d'ordre 24 engendré par a et b ; c'est le groupe des déplacements conservant l'octaèdre régulier dont les sommets sont les centres des huit faces du cube. On peut aussi considérer le sous-groupe cyclique H_1 d'ordre 3 engendré par b : c'est le groupe des rotations qui laissent fixe le sommet 1.

Groupes libres

Si un sous-ensemble K engendre un groupe G, tout élément de G est un produit fini d'éléments de K et d'inverses d'éléments de K, mais l'existence de « relations » entre éléments de K fait qu'il peut y avoir plusieurs telles représentations (cf. supra). Nous allons examiner le cas où il y a unicité.

Soit S un ensemble quelconque. On appelle *mot* de S soit l'ensemble vide, noté ici 1 et appelé le *mot vide*, soit une suite formelle *finie* :

$$a_1 a_2 \dots a_n$$

d'éléments écrits s^{ε} où $s \in$ S, avec $\varepsilon = +1$ ou -1. On dira qu'un tel mot est *réduit* s'il ne contient pas de termes consécutifs de la forme $s^{\varepsilon}s^{-\varepsilon}$, $s \in$ S ; à tout mot, on peut toujours associer le mot réduit obtenu en « gommant », c'est-à-dire en supprimant purement et simplement toute paire consécutive de ce type. On appellera alors *groupe libre* engendré par S l'ensemble des mots réduits muni de la loi de composition qui, à deux mots réduits $f = a_1a_2\ a_n$ et g = $b_1b_2\ b_m$, fait correspondre le mot réduit& obtenu à partir du mot :

$$a_1 a_2 \dots a_n b_1 b_2 \dots b_m,$$

obtenu en écrivant d'abordfpuis g. 11 est clair qu'on obtient bien ainsi un groupe. dont le mot vide est l'élément neutre ; le *mot inverse* est le mot obtenu en renversant l'ordre des termes et en remplaçant l'« exposant » (égal à + 1 ou à $-$ 1) de chaque terme par son opposé : en effet, par réduction on obtient alors le mot vide comme produit d'un mot et du mot inverse puisque, de proche en proche, on « gomme » tout en réduisant le mot obtenu en juxtaposant un mot et son inverse.

Un groupe G est dit *libre* s'il est isomorphe au groupe libre engendré par un ensemble S, que l'on peut supposer inclus dans G en prenant pour S l'ensemble qui correspond dans l'isomorphisme aux mots d'une seule « lettre » avec $\varepsilon = +1$. L'ensemble S est un système de générateurs sans « relations ».

La théorie des groupes libres a été élaborée par J. Nielsen et 0. Schreier. Le résultat principal en est que tout sous-groupe d'un groupe libre est libre (c'est un résultat dont la démonstration est longue et très technique).

3. Relations d'équivalence et quotients

Dans ce qui suit interviendra souvent le fait que l'inverse d'un produit ab de deux éléments d'un groupe est le produit $b^{-1}a^{-1}$

des inverses en renversant l'ordre ; car, en utilisant l'associativité :

$$(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1}$$
$$= a\,1a^{-1} = aa^{-1} = 1.$$

À partir d'un sous-groupe, on peut définir plusieurs relations d'équivalence sur un groupe. Si le sous-groupe vérifie une propriété supplémentaire, ces relations coïncident et on peut alors munir l'ensemble quotient d'une structure de groupe.

Classes suivant un sous-groupe

Soit G un groupe et H un sous-groupe de G. La relation :

$$x \sim_g y \Leftrightarrow x^{-1}y \in \mathrm{H}$$

est une relation d'équivalence sur G. En effet $x \sim_g x$, car $x^{-1}x = 1 \in \mathrm{H}$; si $x \sim_g y$, l'élément $x^{-1}y$ appartient à H et, par suite, aussi son inverse $(x^{-1}y)^{-1} = y^{-1}x$, ce qui signifie : $y \sim_g x$; la transitivité résulte du fait que, si $x^{-1}y$ et $y^{-1}z$ sont deux éléments du sous-groupe H, leur produit $(x^{-1}y)(y^{-1}z) = x^{-1}z$ est aussi un élément de H. La classe d'équivalence d'un élément $x \in$ G est l'ensemble $x\mathrm{H}$ des produits xh lorsque h parcourt H, appelé ***classe à gauche de*** x ***suivant*** H (ou modulo H). Deux classes à gauche sont disjointes ou confondues. Lorsque le nombre de classes à gauche distinctes est fini, on l'appelle ***l'indice*** du sous-groupe H dans G, et on note ce nombre [G : H]. Remarquons que, puisque, pour x fixé, l'application $g \mapsto xg$ est, d'après la règle de simplification (cf. chap. 1), une bijection de G sur lui-même, si le sous-groupe H est fini, toutes les classes à gauche ont le même nombre d'éléments que H. Pour G fini, on obtient, puisque les classes à gauche distinctes forment une partition, que l'ordre du groupe G est le produit de l'ordre de H par l'indice de H dans G, soit :

$$|\mathrm{G}| = [\mathrm{G}:\mathrm{H}]|\mathrm{H}|,$$

résultat obtenu par Lagrange, sous une forme différente, à propos de la théorie des équations et avant l'élaboration de la théorie des groupes proprement dite (cf. la partie C ci-après Groupes finis).

Bien entendu, on pourrait définir de manière analogue les classes à droite $\mathrm{H}x$ pour la relation d'équivalence $xy^{-1} \in \mathrm{H}$. La symétrie $x \mapsto x^{-1}$ est une bijection du groupe sur lui-même qui conserve les sous-groupes et échange les classes à gauche et les classes à droite ; en particulier, si le nombre de classes à gauche suivant H est fini (c'est-à-dire H d'indice fini dans G), ce nombre est aussi le nombre de classes à droite.

Revenons par exemple au groupe des « symétries » du cube, avec les notations du chapitre 2. Le groupe G_1 est d'indice 2 dans G et ici les classes à gauche sont G_1 et $c\,\mathrm{G}_1$, groupe des déplacements directs (rotations) et ensemble des déplacements inverses conservant le cube. Deux éléments x et y de G_1 sont équivalents à gauche par rapport au sous-groupe H' si et seulement s'ils envoient 1 sur le même sommet, puisque les éléments de H_1 conservent ce sommet 1. Il y a donc huit classes : H, $x_2\mathrm{H}$, $x_3\mathrm{H}$, $x_4\mathrm{H}$, $x_5\mathrm{H}$, $x_6\mathrm{H}$, $x_7\mathrm{H}$, $x_8\mathrm{H}$, où x_i est une rotation quelconque de G_1 envoyant le sommet 1 sur le sommet i ; on peut prendre par exemple $x_2 = a$, $x_3 = a'$, $x_4 = a^3$, $x_5 = ba^3$, $x_6 = aba^3$, $x_7 = a^2ba^3$ et $x_8 = a^3ba^3$. Ainsi H' est d'indice 8 dans G_1 et on a bien :

$$[\mathrm{G}_1:\mathrm{H}_1] = \frac{|\mathrm{G}_1|}{|\mathrm{H}_1|}.$$

On voit de même que deux rotations de G_1 sont équivalentes à droite si et seule-

ment si le sommet 1 est l'image du même sommet i par ces deux rotations, ce qui permet d'expliciter les huit classes à droite.

Indiquons enfin que, si H et K sont deux sous-groupes de G, on définit la double classe HxK d'un élément $x \in G$ suivant H et K ; si K $=$ H, on parle de doubles classes suivant H. Comme ci-dessus, on montre, en introduisant une relation d'équivalence convenable, que deux doubles classes sont disjointes ou confondues.

Automorphismes intérieurs

Si G est un groupe, l'ensemble des automorphismes de G est un groupe, que nous noterons Aut(G), pour la composition des applications : c'est le sous-groupe du groupe symétrique C(G) de l'ensemble G, formé des bijections de G sur G qui sont, en plus, des morphismes. Nous allons mettre en évidence certains automorphismes qui jouent un rôle fondamental en théorie des groupes.

Soit s un élément d'un groupe G. L'application a, : G-G définie par $\alpha_s(x) = sxs^{-1}$ est un automorphisme de G que nous appellerons *l'automorphisme intérieur* défini par s ; en effet :

$$\begin{aligned}\alpha_s(xy) &= sxys^{-1} = sx\,1ys^{-1}\\ &= (sxs^{-1})(sys^{-1}) = \alpha_s(x)\alpha_s(y),\end{aligned}$$

et α_s est manifestement bijectif, car $y = \alpha_s(x)$ équivaut à $x = s^{-1}ys$, ce qui montre que l'automorphisme réciproque est l'automorphisme intérieur défini par s^{-1}. De plus, l'application $s \mapsto$ a,, est un morphisme de G dans le groupe Aut(G), car on a :

$$\begin{aligned}\alpha_{ss'}(x) &= ss'x(ss')^{-1}\\ &= ss'xs'^{-1}s^{-1} = s(s'xs'^{-1})s^{-1}\\ &= \alpha_s(\alpha_{s'}(x)) = \alpha_s \circ \alpha_{s'}(x)\,;\end{aligned}$$

par suite, l'ensemble Int(G) de tous les automorphismes intérieurs est un sous-groupe de Aut(G). On appelle *centre* du groupe G le noyau Z(G) du morphisme $s \mapsto$ a, ; c'est un sous-groupe de G qui est l'ensemble des éléments $s \in$ G tels que $sxs^{-1} = x$, soit $sx = xs$, pour tout x $\in$ G. Ainsi le centre est l'ensemble des éléments de G qui *commutent* avec *tous* les éléments du groupe ; ce centre est un sous-groupe commutatif qui est transformé en lui-même par tout automorphisme de G. Plus généralement, soit H un sous-groupe de G et S une partie de G (qui n'est pas nécessairement un sous-groupe). On montre facilement que l'ensemble des éléments $x \in$ H qui commutent avec tous les éléments de S, c'est-à-dire tels que sx = xs pour tout $s \in$ S, est un sous-groupe de H, noté $Z_H(S)$ et appelé le *centralisateur de S dans* H ; avec cette extension, le centre apparaît comme le centralisateur de G dans G.

Soit H un sous-groupe de G. On dit que deux éléments s, s' $\in$ G sont conjugués par rapport à H s'il existe $x \in$ H tel que :

$$s' = xsx^{-1} = \text{a,}(s)$$

il est clair, puisque H est un sous-groupe, que la conjugaison est une relation d'équivalence sur G. Plus généralement, on dit que deux sous-ensembles S et S' de G sont *conjugués par rapport à* H s'il existe $x \in$ H tel que S' $= xSx^{-1}$; si S est un sous-groupe, ses conjugués sont les sous-groupes images de S par les automorphismes intérieurs a_x, $x \in$ H. Revenant à un sous-ensemble S quelconque, on vérifie facilement que l'ensemble $N_H(S)$ des éléments $x \in$ H tels que S = xSx^{-1} est un sous-groupe de H appelé *normalisateur de* S *dans* H ; bien entendu, $Z_H(S)$ est un sous-groupe de $N_H(S)$.

Dans tout ce qui précède, on supprime la référence à H si H = G ; on parle alors d'éléments conjugués, de centralisateur, de normalisateur.

Sous-groupes distingués

Si un groupe H est le noyau d'un morphisme f d'un groupe G dans un groupe G', pour tout $x \in$ G et y $\in$ H, donc $f(y) = 1$, on a :

$$f(xyx^{-1}) = f(x)f(y)f(x^{-1}) = f(x)f(x^{-1}) = f(xx^{-1}) = 1,$$

d'après (2) et, par suite, $xyx^{-1} = \alpha_x(y) \in$ H ; ainsi xHx^{-1} = H pour tout $x \in$ G et H est égal à tous ses conjugués. On dit qu'un sous-groupe possédant cette propriété est *distingué (ou normai, ou invariant)* ; ainsi les noyaux des morphismes sont des sous-groupes distingués. En fait, on va voir aussi que tout sous-groupe distingué est le noyau d'un certain morphisme.

Revenons pour un instant à un sous-groupe quelconque H d'un groupe G et désignons par G/H l'ensemble des classes à gauche suivant H, c'est-à-dire l'ensemble quotient de G par la relation d'équivalence à gauche suivant H, et par f : G $\rightarrow$ G/H l'application canonique qui à tout x $\in$ G associe sa classe à gauche. S'il est possible de munir G/H d'une structure de groupe pour laquelle f est un morphisme, alors, d'après (2), l'élément neutre de ce groupe est la classe de 1 et le noyau de f est donc H ; ainsi H est nécessairement un sous-groupe distingué. Réciproquement, si on suppose maintenant H distingué, la classe d'un produit xy ne dépend que des classes de x et y, car, si $x^{-1}x' \in$ H et $y^{-1}y' \in$ H, c'est-à-dire x et y équivalents (à gauche) à x' et y' respectivement, on a :

$$(xy)^{-1}x'y' = y^{-1}x^{-1}x'y' = (y^{-1}x^{-1}x'y)(y^{-1}y') \in H,$$

comme produit de deux éléments de H (le premier est l'image de $x^{-1}x'$ par l'automorphisme intérieur défini par y), et on vérifie facilement qu'on peut ainsi munir G/H d'une structure de groupe pour laquelle l'application canonique est un morphisme. Le groupe G/H est appelé le *groupe quotient* de G par le sous-groupe distingué H ; si H n'est pas distingué, on peut cependant faire « opérer » le groupe sur l'espace G/H et on obtient alors ce qu'on appelle un espace homogène (cf. chap. 5). L'application canonique f réalise une bijection entre les sous-groupes de G contenant H et les sous-groupes de G/H, à un sous-groupe distingué correspondant un sous-groupe distingué et vice versa.

Si f : G $\rightarrow$ G' est un morphisme de noyau H, le premier théorème d'isomorphisme des groupes affirme que le groupe quotient G/H est isomorphe au groupe $f(G)$ image de G par f par le morphisme qui, à chaque classe, fait correspondre la valeur constante de f sur cette classc. Il cn résulte que f peut s'écrire comme composé de trois morphismes qui sont (respectivement de gauche à droite) surjectif, bijectif et injectif :

$$G \rightarrow G/H \text{ -} f(G) \rightarrow G$$

Indiquons enfin un autre résultat d'isomorphisme. Si H est un sous-groupe distingué de G et L un sous-groupe quelconque, alors LH = HL est un sous-groupe, H $\cap$ L est un sous-groupe distingué de L et les deux groupes L/(H $\cap$ L) et HL/H sont isomorphes,

Suites de composition

Dans un groupe G, le sous-groupe {1} réduit à l'élément neutre et le groupe G lui-même sont distingués ; si ce sont les seuls sous-groupes distingués de G, ce groupe est dit *simple.* À l'opposé, dans un groupe commutatif, tous les sous-groupes sont distingués. Nous allons expliquer maintenant comment on peut préciser la structure d'un groupe en fabriquant des

suites de sous-groupes encastrés. Pour éviter des interprétations erronées, il sera bon de se rappeler qu'il n'y a pas transitivité de la notion de sous-groupe distingué : si K C H C G, K sous-groupe distingué de H et H sous-groupe distingué de G, le sous-groupe K de G n'est pas nécessairement distingué dans G.

On appelle *suite de composition* d'un groupe G une suite strictement décroissante :

$$G_0 = G \supset G_1 \supset G_2 \supset ... \supset G_n = \{1\}$$

de sous-groupes telle que G_{i+1} soit un sous-groupe *distingué* de G_i pour $i = 0$, 1, ... $n-1$; les groupes quotients G_i/G_{i+1} s'appellent les *facteurs de composition* de la suite ; si tous les facteurs de composition sont simples, on dit qu'on a une suite de Jordan-Holder ; dans un groupe fini, on peut toujours trouver de telles suites de Jordan-Holder et leur longueur est un important invariant du groupe (cf. théorème de Jordan-Holder dans la partie C ci-après Groupes finis, chap. 2).

Examinons ici les groupes dits *résolubles* admettant des suites de composition dont tous les facteurs de composition sont commutatifs ; historiquement, cette notion est liée à la résolubilité des équations algébriques par radicaux, d'où la terminologie (cf. CORPS, chap. 3). Nous aurons pour cela besoin d'une condition exprimant que le quotient G/H d'un groupe G par un sous-groupe distingué H est commutatif. S'il en est ainsi, quels que soient x et y dans G, les classes de xy et yx suivant H doivent être égales ; donc l'élément :

$$(x,y) = (xy)^{-1}yx = y^{-1}x^{-1}yx,$$

appelé *commutateur* de x et y, doit appartenir à H. De manière générale, si H et K sont des sous-groupes de G, notons (H, K) le groupe *engendré* par les commutateurs (x, y) pour $x \in H$ et $y \in K$ (attention, le produit de deux commutateurs n'est pas un commutateur en général). Le groupe quotient G/H est donc commutatif si et seulement *si* H contient le groupe (G, G) = D(G) appelé *groupe des commutateurs de G,* ou *groupe dérivé* ; ce groupe est évidemment transformé en lui-même par tout automorphisme de G et, en particulier, il est distingué dans G. On peut donc itérer cette opération de « dérivation » et construire la suite décroissante des groupes dérivés successifs :

$$G \supset D(G) \supset D^2(G) = D(D(G)) \supset . \supset D^{r+1}(G) = D^r(D(G)) ;$$

on montre alors qu'un groupe G est résoluble si et seulement s'il existe un entier r tel que $D^{r+1}(G) = \{1\}$ et qu'alors il existe une suite de composition dont tous les termes G_i sont des sous-groupes distingués dans G (et non pas seulement dans le sous-groupe précédent G_{i-1}) et dont tous les facteurs de composition sont commutatifs.

Si G est un groupe *fini*, il est résoluble si et seulement s'il admet une suite de Jordan-Holder dont tous les facteurs de composition sont des groupes cycliques d'ordre premier.

On montre que tout sous-groupe ou tout groupe quotient d'un groupe résoluble est résoluble.

Un cas particulier de groupes résolubles est fourni par les groupes *nilpotents* qui sont les groupes admettant une suite de composition :

$$G \supset G_1 \supset G_2 \supset ... \supset G_n = \{1\}$$

telle que chaque groupe quotient G_{i-1}/G_i soit dans le centre du groupe G/G_i (une telle suite est dite centrale) ; dans le cas de groupes finis (cf. la partie C ci-après - Groupes finis, chap. 3), on peut donner

diverses autres caractérisations, Définissons par récurrence le commutateur de k éléments $x_1, \ldots, x_k$ de G comme :

$$(x_1, x_2, \ldots, x_k) = ((x_1, \ldots, x_{k-1}), x_k)$$

et désignons par $C^k(G)$ le groupe engendré par les commutateurs de k éléments de G. On vérifie que les groupes $C^k(G)$ peuvent être définis par les relations de récurrence :

$$C^2(G) = D(G), \quad C^{r+1}(G) = (C^r(G), G),$$

et forment une suite décroissante de sous-groupes (appelée série centrale descendante). Un groupe G est alors nilpotent si et seulement s'il existe un entier r tel que $C^{r+1}(G) = \{1\}$.

4. Produits

Soit n groupes $G_1, \ldots, G_n$. L'ensemble produit :

$$G = G_1 \times G_2 \times \ldots \times G_n$$

est un groupe, appelé *groupe produit,* pour la loi de composition :

$$(x_1, x_2, \ldots, x_n)(y_1, y_2, \ldots, y_n) = (x_1 y_1, x_2 y_2, \ldots, x_n y_n);$$

si $H_1, H_2, \ldots, H_n$ sont des sous-groupes de G_1, G_2, ..., G_n respectivement, le groupe produit :

$$H_1 \times H_2 \times \ldots \times H_n$$

est un sous-groupe de G, distingué si chacun des H_i l'est. Prenons en particulier $H_i = G_i$ et $H_j = \{1\}$ pour $j \neq i$; le groupe produit est un sous-groupe distingué de G isomorphe à G_i et nous identifierons ces deux groupes. Remarquons que, en effectuant cette identification, tout élément de G_i commute avec tout élément de G_j pour $i \neq j$.

Dans la situation précédente, le groupe G apparaît comme produit de certains de ses sous-groupes et les groupes obtenus en changeant l'ordre des groupes facteurs sont isomorphes.

On dira qu'un groupe G est *produit direct* d'une famille finie $H_1, \ldots, H_n$ de sous-groupes distincts de G si tout élément de H_i commute avec tout élément de H_j pour $i \neq j$ et si tout élément u de G s'écrit *de manière unique* comme un produit :

$$u = u_1 u_2 \ldots u_n, \quad u_i \in H_i;$$

on dit que u_i est le composant de u dans H_i. Cela entraîne que les H_i sont des sous-groupes distingués de G et que l'application :

$$(u_1, u_2, \ldots, u_n) \mapsto u_1 u_2 \ldots u_n$$

est un isomorphisme du groupe produit $H_1 \times \ldots \times H_n$ sur le groupe G. Si $H_1, H_2, \ldots, H_n$ sont des sous-groupes distingués d'un groupe G tels que :

$$(H_1 H_2 \ldots H_i) \cap H_{i+1} = \{1\}, \; i = 1, \ldots, n-1,$$

on montre que l'ensemble $H_1 H_2 \ldots H_n$ est un sous-groupe distingué de G qui est produit direct de la famille considérée.

Un sous-groupe distingué H d'un groupe G est dit *facteur direct* dans G s'il existe un sous-groupe distingué K de G tel que G soit égal au produit direct de H et K ; remarquons que le groupe K est alors isomorphe au groupe quotient G/H.

En notation additive, on parle de somme directe au lieu de produit direct.

Le produit direct permet de définir une nouvelle et importante classe de groupes (cf. ci-après les parties B – Groupes classiques et géométrie et E Groupes de Lie). Un groupe G est dit *semi-simple* s'il est produit direct d'un nombre fini de sous-groupes simples (c'est-à-dire dont les seuls sous-groupes distingués sont tri-

viaux). On montre que le nombre de ces sous-groupes, appelé la *longueur* de G, est le même pour toutes les expressions de G comme produit direct de sous-groupes simples. Si G est produit direct d'une famille finie (H,), $i \in I$, de sous-groupes simples, tout sous-groupe distingué K est isomorphe au produit direct d'une sous-famille (H,), $j \in J$, J C 1 ; en particulier tout sous-groupe distingué est semi-simple, de longueur inférieure ou égale à celle de G (avec égalité des longueurs si et seulement si $K = G$). Il en est de même des groupes quotients d'un groupe semi-simple.

On va maintenant généraliser la notion de produit direct. Soit H et K deux groupes, et soit donné, pour tout $x \in H$, $x \mapsto \tau_x$ un morphisme de H dans le groupe Aut (K) des automorphismes de K. On appelle *produit semi-direct de* H *pur* K *relatif à* τ l'ensemble H X K, muni de la loi de composition :

$$(x, y)(x', y') = (xx', \tau_{x'}(y)y'),$$

qui est un groupe ; on notera $H \times_\tau K$ ce produit semi-direct. Si $\tau_x(y) = y$ pour tout $x \in H$, on retrouve le produit direct défini ci-dessus. On vérifie que les éléments de la forme $(x, 1)$, $x \in H$, forment un sous-groupe de G isomorphe à H et que les éléments de la forme $(1, y)$, $y \in K$, forment un sous-groupe distingué de G isomorphe à K. Réciproquement, soit G un groupe, H un sous-groupe de G et K un sous-groupe distingué de G tel que $H \cap K = \{1\}$. Cela a pour conséquence que $xy = x'y'$, pour $x, x' \in H$ et y, $y' \in K$, entraîne $x = x'$ et $y = y'$ (car on a alors $x'^{-1}x = y'y^{-1} \in H \cap K$, d'où $x'^{-1}x = y'y^{-1} = 1$). Puisque K est distingué, pour tout $x \in H$, l'automorphisme intérieur défini par x^{-1} induit un automorphisme de K que nous désignerons par τ_x. On voit alors facilement que HK est un groupe isomorphe au produit semi-direct de H par K relativement à τ ; en effet, pour $x, x' \in H$ et y, $y' \in K$, on a :

$$xyx'y' = xx'(x'^{-1}yx')y' = xx'\tau_{x'}(y)y' ;$$

le groupe quotient HK/K est isomorphe à H.

5. Groupes de transformations

Si E est un ensemble, nous avons déjà indiqué que les bijections de E sur lui-même forment un groupe Σ (E) pour la composition des applications, le *groupe symétrique* de E. Si E est muni d'une structure, les bijections qui conservent cette structure forment un sous-groupe de C(E), *le groupe des automorphismes de* E pour la structure considérée. C'est ainsi qu'on a introduit ci-dessus le groupe Aut(G) des automorphismes d'un groupe G ; si V est un espace vectoriel, on obtient *le groupe linéaire* de V, noté GL(V), formé des bijections linéaires de V sur V.

On dit qu'un groupe G *opère* sur un ensemble E si E est muni d'une loi externe dont le domaine d'opérateurs est G :

$$(g, x) \mapsto gx,$$

de telle sorte que $g(hx) = (gh)x$ et $1x = x$ pour g, $h \in G$ et $x \in E$. Cela entraîne que, pour $g \in G$, l'application p(g) : E $\to$ E qui à x fait correspondre gx est une bijection de E sur lui-même (dont la bijection réciproque est $\rho(g^{-1})$) ; la condition d'associativité s'écrit $\rho(gh) = p(g) \circ p(h)$ et exprime donc que ρ est un morphisme de G dans le groupe symétrique C(E). On appelle un tel morphisme une *représentation* du groupe G dans le groupe $\Sigma(E)$; si ρ est un isomorphisme de G sur son image,

on dit qu'on a réalisé le groupe G comme groupe de transformations de E. Remarquons qu'on peut toujours réaliser un groupe comme groupe de transformations de l'ensemble qui lui est sous-jacent en identifiant tout élément $g \in G$ à la translation à gauche $h \mapsto gh$. Dans ce qui suit, nous considérerons un groupe G qui opère sur un ensemble E.

Pour $x \in E$, on appelle orbite de x l'ensemble des éléments gx pour $g \in G$; remarquons que les orbites de deux éléments sont toujours disjointes ou confondues, car la relation $x \sim y$, s'il existe $g \in G$ tel que $y = gx$, est une relation d'équivalence sur E. On appelle *classes d'intransitivité* les classes pour cette relation d'équivalence, c'est-à-dire les orbites disjointes. Le groupe est dit *transitif* s'il n'existe qu'une seule classe d'intransitivité (E tout entier). Cela signifie que, si x et y sont deux éléments quelconques de E, il existe au moins un élément $g \in G$ tel que $y = gx$; si cet élément g est de plus toujours unique, le groupe est dit *simplement transitif*. Plus généralement, on dit que le groupe G est *n-fois transitif* si E contient au moins n éléments et si, étant donné deux systèmes quelconques $x_1, ..., x_n$ et $y_1, ..., y_n$ de n éléments de E, il existe au moins un élément $g \in G$ tel que $y_i = gx_i$ pour $i = 1, 2, ..., n$. On verra de nombreux exemples de ces situations dans les articles sur les groupes classiques et sur les groupes finis.

On appelle enfin *espace homogène* un ensemble E muni d'un groupe transitif d'opérateurs. Voici, pour terminer, un exemple important de cette situation, auquel on peut toujours se ramener par un isomorphisme. Soit G un groupe et H un sous-groupe quelconque de G ; désignons par G/H l'ensemble des classes à gauche suivant H. Il est clair que, pour $g \in G$, l'application qui à la classe à gauche de x fait correspondre la classe à gauche de gx est une bijection de G/H sur lui-même et que l'on fait ainsi opérer G transitivement sur l'ensemble G/H, ce qui munit cet ensemble d'une structure d'espace homogène. Réciproquement, si E est un ensemble sur lequel opère transitivement un groupe G, soit a un élément de E et désignons par H, le sous-groupe des éléments de G laissant a invariant, c'est-à-dire tels que $ga = a$; on vérifie facilement que l'espace homogène E est isomorphe (en tant qu'espace homogène) à l'espace homogène G/H_a des classes à gauche de G suivant le sous-groupe H_a.

JEAN-LUC VERLEY

Bibliographie

J. CALAIS, *Éléments de théorie des groupes*, P.U.F., Paris, 1984 / M. KARGAPOLOV & J. MERZLJAKOV, *Éléments de théorie des groupes*, Mir, Moscou, 1985 / A. G. RUROSH, *Theory*, 2 vol., Chelsea Publ., New York, 1979 / E. SCHENKMAN, *Group Theory*, repr. of 1965, Krieger Publ., Melbourne (Fla.), 1975 / W. R. SCOTT, *Group Theory*, repr. of 1964, Dover Publ, New York, 1987.

B. Groupes classiques et géométrie

Jusque vers 1800, la géométrie dite « élémentaire » est restée à peu de chose près ce qu'elle était dans l'Antiquité, tant dans sa substance que dans ses méthodes (l'invention de la « géométrie analytique » ayant à peu près exclusivement servi à prolonger le champ d'action de la géométrie classique dans les directions de la géométrie algébrique et de la géométrie différentielle). Mais, même dans les exposés d'Euclide et de ses continuateurs, bien que l'intérêt se concentre sur les propriétés des figures « classiques » (triangle, rectangle, parallélogramme, cercle, coniques,

etc.), les *isométries* (transformations de l'espace ou du plan conservant les distances) jouent un rôle essentiel, non toujours explicité ; le fait qu'elles forment un groupe était implicitement utilisé bien avant que la notion abstraite de groupe ne se fût dégagée. À partir de 1800 environ. avec le développement de la géométrie projective, on commence à distinguer, parmi les notions géométriques invariantes par isométrie, celles qui sont de nature « descriptive » de celles que l'on qualifie de « métriques », les premières restant invariantes par des transformations *plus générales*, à savoir celles qui transforment *linéairement* les coordonnées cartésiennes ; par exemple, dans le plan, au point (x, y) correspond le point (x', y') tel que :

$$x' = ax + bx + c, \quad y' = a'x + b'y + c'.$$

C'est ainsi que, par une telle transformation, une médiane d'un triangle se transforme en une médiane du triangle image : la notion de médiane est « descriptive » ; au contraire, une hauteur d'un triangle n'a pas cette propriété : la notion de hauteur est « métrique ». Avec Félix Klein et son ((programme d'Erlangen » (1872), cette distinction s'est précisée, et le concept même de « géométrie » a reçu une définition générale, englobant la géométrie classique (dite aussi « euclidienne »), la géométrie projective, la géométrie conforme, les géométries « non euclidiennes », etc. : une géométrie est l'étude des notions et des propriétés qui restent *invariantes par un groupe donné de transformations*. De ce fait, la « géométrie », après Klein, est devenue essentiellement l'étude de ces groupes, les propriétés des « figures » classiques passant au second plan ; plus généralement, dans toutes les parties des mathématiques où intervient un *espace homogène* G/H (ou, ce qui revient au même, un espace dans lequel un groupe G opère *transitivement)*, un principe fécond est d'en ramener l'étude à celle du groupe G lui-même.

Les groupes envisagés par Klein et certaines de leurs généralisations sont connus sous le nom de « groupes classiques » ; en tant que groupes de Lie, ils correspondent aux algèbres de Lie *simples* « classiques » (cf. la partie E ci-après-Groupes de Lie) et, de ce fait, la théorie des représentations linéaires (de dimension finie) et des *invariants* de ces groupes peut être regardée comme entièrement connue *(ibid.)* ; ce qui, en un certain sens, permet de considérer les « géométries » correspondantes comme essentiellement achevées et ne présentant plus aucun problème digne de recherches mathématiques sérieuses.

Nous allons parler d'abord en détail des deux groupes les plus liés à la géométrie classique, le *groupe linéaire général* et le *groupe orthogonal* ; mais nous nous placerons d'emblée dans la géométrie à n dimensions (n arbitraire $\geqslant 2$). Nous supposons connus du lecteur les notions et résultats fondamentaux de l'algèbre linéaire et multilinéaire, exprimés dans le langage géométrique des espaces vectoriels ou projectifs ; il pourra voir combien l'algèbre linéaire facilite, dans ces groupes, la solution de problèmes qui présentent de grandes difficultés dans des groupes quelconques.

1. Le groupe linéaire général

Soit E un espace vectoriel de dimension n sur le corps R des nombres réels ; on appelle *groupe linéaire général* de E et on note GL(E) le groupe de tous les automorphismes de l'espace vectoriel E (ou transformations linéaires de E en lui-

même) ; il est isomorphe au groupe $GL(n, R)$ des *matrices inversibles* d'ordre n sur R. L'application $u \mapsto \det(u)$, où $\det(u)$ désigne le déterminant de u, est un homomorphisme de GL(E) sur le groupe multiplicatif R * des nombres réels $\neq 0$; le noyau SL(E), ou $SL(n, R)$, de cet homomorphisme est appelé *groupe unimodulaire ou groupe linéaire spécial.*

Générateurs

On caractérise aisément les *involutions* de GL(E), transformations u telles que $u^2 = 1$ ou $u^{-1} = u$. Comme on peut écrire :

$$x = \frac{1}{2}(x + u(x)) + \frac{1}{2}(x - u(x)),$$

on voit que E est somme directe de deux sous-espaces V^+, V dans lesquels on a respectivement u(x) = x et $u(x) = -x$ (ce sont donc les sous-espaces propres de u pour les valeurs propres 1 et -1) ; on a :

$$\det(u) = (-1)^{\dim V^-},$$

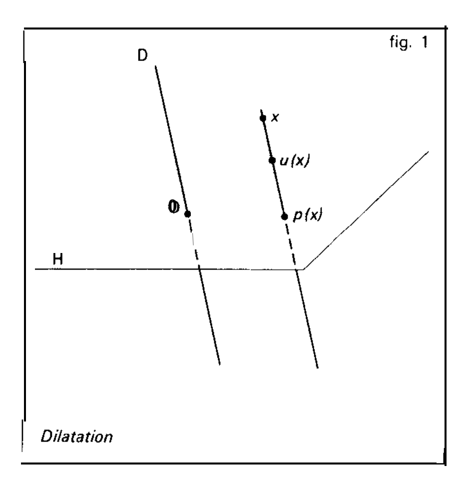
où dim V^- est la dimension de l'espace vectoriel V^-. Si $V^- = \{0\}$, u est l'identité ; si $V^+ = \{0\}$, u est la symétrie $x \mapsto -x$. Lorsque V^+ est un hyperplan H, on dit que u est une *réflexion* d'hyperplan H.

Si H est un hyperplan d'équation f(s) = 0 (f forme linéaire), les transformations de GL(E) qui laissent invariants tous les points de H sont de deux sortes :

a) les *dilatations* : une telle transformation u est définie par une droite D supplémentaire de H et par un nombre $\lambda \neq 1$ tels que $u(x) = \lambda x$ dans D, d'où :

$$u(x) = p(x) + \lambda(x - p(x));$$

pour x quelconque dans E, *p(x)* étant la projection de x sur H parallèlement à D



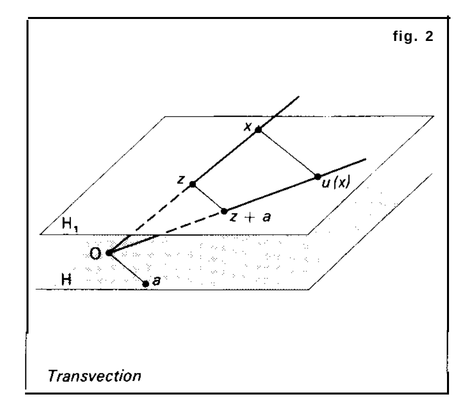
Dilatation

(fig. 1). On a $\det(u) = \lambda$; pour $\lambda = -1$, on obtient les réflexions d'hyperplan H.

b) les transvections, de la forme :

$$u(x) = x + f(x)a,$$

ou $a \in$ H ; dans l'hyperplan affine H, parallèle à H, d'équation $f(x) = 1$, u est la translation $z \mapsto z + a$ (fig. 2).



Transvection

On dit que u est une transvection d'hyperplan H et de droite D = Ra ; on a $\det(u) = 1$. Les transvections forment un système générateur de SL(E) ; les dilatations et les transvections engendrent GL(E). Toute transvection est produit de deux réflexions.

Centralisateurs

Soit Z $=$ Z(E) le groupe des homothéties $h_\lambda : x \mapsto$ Ax de rapport A $\neq 0$, qui est isomorphe à R *. Une homothétie peut être caractérisée comme une transformation de GL(E) laissant invariante (globalement) toute droite de E.

Pour tout hyperplan H de E, les transvections d'hyperplan H forment un groupe commutatif Θ(E, H) isomorphe au groupe additif H ; son centralisateur dans GL(E) est le groupe ZO(E, H). Une transvection $\neq 1$ et une dilatation ne sont jamais permutables ; le centralisateur dans GL(E) du sous-groupe Γ(E, H) laissant invariants tous les points de H est donc réduit à Z ; en particulier Z est le *centre* de GL(E), et Z $\cap$ SL(E) le centre de SL(E) (isomorphe au sous-groupe de $\lambda \in$ R tels que $\lambda^n = 1$, donc réduit à l'élément neutre, si n est impair, et formé de l'identité et de la symétrie $x \mapsto -x$, si n est pair).

Les transvections de droite donnée D (et d'hyperplan variable contenant D) forment un groupe commutatif Θ' (E, D) ; le centralisateur de ce sous-groupe dans GL(E) est $Z\Theta'$(E, D).

Propriétés de transitivité et de conjugaison

Le groupe SL(E), et a fortiori GL(E), opère de façon *doublement transitive* sur les droites de E ; si (D_1, D_2) et (D'_1, D'_2) sont deux couples de droites distinctes, il existe au moins une transformation $u \in$ SL(E) telle que $u(D_1) = D'_1$, $u(D_2) = D'_2$; si $n = 2$, GL(E) opère de façon *triplement transitive* sur les droites de E. En général, on appelle *repère projectif* de E un système (D_j) de $n + 1$ droites de E dont n quelconques ne sont pas dans un même hyperplan ; pour deux repères projectifs (D,) et (D',), il existe $u \in$ GL(E) tel que $u(D_j) = D'$, pour $1 \leqslant j \leqslant n+1$, et u est déterminé à un facteur h_λ près.

Deux réflexions quelconques sont conjuguées dans GL(E) ; ii en est de même de deux transvections $\neq 1$. Si $n \geqslant 3$, deux transvections $\neq 1$ sont conjuguées dans SL(E) ; au contraire, si $n = 2$, il y a deux classes de transvections conjuguées dans SL(E) : si t est une transvection $\neq 1$, t et t^k sont conjuguées dans SL(E) pour $k > 0$ mais non pour $k < 0$.

En tout cas, comme :

$$t = t^2 t^{-1} = (sts^{-1})t^{-1},$$

pour un $s \in$ SL(E), toute transvection est un *commutateur* de deux éléments de SL(E) et, par suite, SL(E) est égal à son *groupe des commutateurs*, qui est aussi le groupe des commutateurs de GL(E).

Simplicité du groupe SL(E)/(Z $\cap$ SL(E))

On va voir que tout sous-groupe *distingué* N de SL(E) est soit contenu dans le centre Z $\cap$ SL(E), soit égal à SL(E). Supposons donc N $\not\subset$ Z.

a) *N* opère *transitivement* sur les droites de E. En effet, si $u \in$ N n'est pas dans Z, il existe au moins une droite D telle que u(D) $\neq$ D ; pour toute autre droite D', il y a un v $\in$ SL(E) tel que v(D) = D', donc $uv^{-1}(D') \neq v^{-1}(D')$ ou $vuv^{-1}(D') \neq D'$, et on a vuv^{-1} EN ; donc, pour *toute* droite D_1 de E, il y a un $u_1 \in$ N tel que $u_1(D_1) \neq D_1$. Soit D_2 une droite distincte de D_1 et montrons qu'il existe $u \in$ N tel que $u(D_1) = D_2$. Comme SL(E) opère de façon doublement transitive sur les droites de E, il existe $v \in$ SL(E) tel que $v(D_1) = D_1$, $v(D_2) = u_1(D_1)$; alors $v^{-1}u_1v(D_1) = D_2$ et $v^{-1}u_1v \in$ N.

b) Pour toute droite D de E, soit S_D l'ensemble des v $\in$ SL(E) tels que

$v(D) = D$; alors $SL(E) = N . S_D$; en effet, pour tout $u \in SL(E)$, il existe $v \in N$ tel que $u(D) = v(D)$, donc $v^{-1} u \in S_D$.

c) S_D contient le groupe commutatif $\Theta'(E, D)$, qui est évidemment distingué dans S_D, et on a vu (cf. *Générateurs, supra)* que les conjugués dans $SL(E)$ du groupe $\Theta'(E, D)$ engendrent $SL(E)$. Tout $v \in SL(E)$ peut donc s'écrire :

$$\prod s_i t_i s_i^{-1}$$

avec $t_i \in \Theta'(E,D)$; puis on peut écrire $s_i = u_i v_i$ avec $u_i \in N$ et $v_i \in S_D$, en vertu du chapitre 2 ; puisque $\Theta'(E, D)$ est distingué dans S_D, on a :

$$s_i t_i s_i^{-1} = u_i t'_i u_i^{-1},$$

avec $t'_i \in \Theta'(E,D)$. Mais, comme N est distingué dans SL(E), un produit d'éléments de $\Theta'(E, D)$ et d'éléments de N (dans n'importe quel ordre) peut toujours s'écrire ut avec $u \in N$ et $t \in \Theta'(E, D)$. Considérons alors deux éléments $v_1 = u_1 t_1$, $v_2 = u_2 t_2$ de SL(E), avec u_1, u_2 dans N et t_1, t_2 dans $\Theta'(E, D)$; on a donc :

$$\begin{aligned} v_1 v_2 v_1^{-1} v_2^{-1} &= u_1 (t_1 u_2 t_1^{-1}) t_1 t_2 t_1^{-1} t_2^{-1} (t_2 u_2^{-1} t_2^{-1}) u_2^{-1} \\ &= u_1 (t_1 u_2 t_1^{-1})(t_2 u_1^{-1} t_2^{-1}) u_2^{-1} \in N, \end{aligned}$$

en vertu de la *commutativité* du groupe $\Theta'(E, D)$. Mais on a vu plus haut (*Propriétés de transitivité et de conjugaison)* que $SL(E)$ est égal à son groupe des commutateurs ; donc $N = SL(E)$.

2. Le groupe orthogonal

On suppose donné sur E un *produit scalaire* : c'est une application *bilinéaire :*

$$(x, y) \mapsto (x \mid y)$$

de $E \times E$ dans R, qui est en outre supposée *symétrique*, c'est-à-dire que :

$$(x \mid y) = (y \mid x),$$

et *positive* non *dégénérée,* c'est-à-dire que :

$$(x \mid x) > 0$$

pour $x \neq 0$ dans E. La donnée d'une telle application définit dans E une notion *d'orthogonalité* ; x, y dans E sont dits *orthogonaux* si l'on a $(x \mid y) = 0$ (relation *symétrique* en x et y). On dit que deux sous-espaces vectoriels V, W de E sont *orthogonaux* si tout vecteur de V est orthogonal à tout vecteur de W ; pour un sous-espace vectoriel V donné, l'ensemble des vecteurs orthogonaux à tous les vecteurs de V est le plus grand sous-espace vectoriel orthogonal à V ; on l'appelle *l'orthogonal* de V et on le note $V^{\perp}$. On a les relations :

$$\begin{gathered} V \cap V^{\perp} = \{0\}, \\ V + V^{\perp} = E, \\ \dim V + \dim V^{\perp} = \dim E, \\ (V + W)^{\perp} = V^{\perp} \cap W^{\perp}, \\ (V \cap W)^{\perp} = V^{\perp} + W^{\perp}, \\ (V^{\perp})^{\perp} = V. \end{gathered}$$

L'exemple classique de produit scalaire dans $\mathbf{R}^n$ est :

$$(x \mid y) = \sum_{j=1}^{n} \xi_j \eta_j ;$$

inversement, pour tout produit scalaire $(x \mid y)$ sur E, il existe une base dite *orthonormale* (e_j) de E telle que :

$$\begin{aligned} (e_i \mid e_j) &= 0 \quad \text{si } i \neq j, \\ (e_i \mid e_i) &= 1 \quad \text{pour tout } i. \end{aligned}$$

Un espace vectoriel E muni d'un produit scalaire est ce qu'on appelle un espace *euclidien* ; sur un même espace vectoriel E, il y a une infinité de produits scalaires non

proportionnels, donnant une infinité de structures d'espace euclidien pour lesquelles les notions d'orthogonalité sont distinctes ; toutefois tous ces espaces sont isomorphes, en vertu de l'existence des bases orthonormales. On suppose dans ce qui suit que le produit scalaire est *fixé*, et on pose $\|x\| = (x \mid x)^{1/2}$ (*longueur* du vecteur x).

On appelle *similitude* de E une transformation linéaire $u \in GL(E)$ telle que :

$$(u(x) \mid u(y)) = \mu(x \mid y),$$

quels que soient x, y dans E, où $\mu = \mu(u)$ est une constante $\neq 0$ dite *multiplicateur* de u ; on a nécessairement $\mu > 0$ comme on le voit en faisant $y = x \neq 0$. Si U est la matrice de u rapporté à une base *orthonormale,* il revient au même de dire que :

$$(1) \qquad {}^tU = \mu U^{-1}.$$

Les similitudes forment un sous-groupe GO(E) C GL(E), et $u \mapsto \mu(u)$ est un homomorphisme de ce groupe sur le groupe multiplicatif RT des nombres réels > 0 ; son noyau O(E) est appelé le *groupe orthogonal* de E (pour le produit scalaire considéré) ; c'est donc le sous-groupe de GL(E) formé des u tels que :

$$(u(x) \mid u(y)) = (x \mid y) ;$$

on peut montrer que c'est aussi le groupe de toutes les applications non supposées linéaires a priori telles que $u(0) = 0$, $\| u(x) \| = \| x \|$ pour tout $x \in E$.

Toute homothétie h_λ est une similitude de multiplicateur λ^2 ; toute similitude de multiplicateur μ s'écrit d'une seule manière $h_\lambda \cdot v$, où $\lambda = \sqrt{\mu}$ et $v \in O(E)$; GO(E) est produit direct du groupe O(E) et du groupe multiplicatif Z+(E) des homothéties de rapport > 0, isomorphe à RT.

Pour une transformation orthogonale de matrice U, on a, d'après la formule (1), $(\det U)^2 = 1$; le sous-groupe O+(E), ou SO(E), des transformations orthogonales de déterminant 1 (aussi appelées *rotations)* est d'indice 2 dans O(E). Les similitudes appartenant au sous-groupe :

GO+(E) = O+(E) X Z+(E)

sont dites *directes,* les autres *inverses.*

Lorsque $E = \mathbf{R}^n$, on suppose toujours que $\mathbf{R}^n$ est muni du produit scalaire classique, et on écrit O(n,-R) [resp. $O^+(n, R)$ ou $SO(n, R)$] au lieu de O(R") [resp. O+(R")] et on l'identifie avec le groupe des matrices orthogonales (*i.e.* telles que ${}^tU = U^{-1}$). Si E est de dimension n, le groupe O(E) est isomorphe à O(n, R).

Générateurs du groupe orthogonal

Les *involutions* u de GL(E) qui appartiennent à O(E) sont celles pour lesquelles les sous-espèces propres V^+ et V^- (cf. *Générateurs,* in chap. 1) sont *orthogonaux :* on dit encore qu'une telle involution est une *symétrie orthogonale* par rapport à V^+. Lorsque V^+ est un hyperplan H, on dit encore *réflexion orthogonale* de droite $V^- = H^\perp$. Si dim $E = n$, toute transformation orthogonale est produit de n réflexions orthogonales au plus. Lorsque V^+ est de dimension $n - 2$, on dit que l'involution est un *renversement d'axe* V^- ; pour $n \geqslant 3$, toute rotation est produit de n renversements au plus. Tout renversement est un *commutateur* de O+(E) si $n \geqslant 3$: en effet, soit (e_1, e_2) une base orthonormale de V^-, et posons $V^+ = \mathbf{R}e_3 \oplus W$, où W est orthogonal à e_3 ; on peut écrire $u = v_1 v_2$, où v_1 (resp. v_2) est le renversement d'axe $\mathbf{R}e_1 \oplus \mathbf{R}e_3$ (resp. $\mathbf{R}e_2 \oplus \mathbf{R}e_3$) ; comme v_2 est conjugué de v_1 dans O^+ (E) [cf. *infra, Propriétés de transitivité et de conjugaison]* et comme $v_1 =$

v_1^{-1}, on a $u = v_1^{-1} s v_1 s^{-1}$ pour un $s \in O'(E)$. On en conclut que O+(E) est son propre groupe des commutateurs et le groupe des commutateurs de O(E).

Le centre Z_0 de O(E) est formé de l'identité et de la symétrie $x \mapsto x$. Si n est pair, Z_0 est aussi le centre de $O^+(E)$; sinon, le centre de O+(E) est réduit à l'identité et O(E) est le produit direct $Z_0 \times O^+(E)$.

Propriétés de transitivité et de conjugaison

Pour que deux sous-espaces vectoriels V_1, V_2 de E soient transformés l'un de l'autre par une transformation orthogonale, il faut et il suffit qu'ils aient même dimension ; il existe alors une rotation u telle que $V_2 = u(V_1)$. Les symétries orthogonales par rapport à V_1 et V_2 sont alors conjuguées.

Le groupe O(2, R) et les angles

Pour une matrice U d'ordre 2, le calcul montre que la relation (1) équivaut à dire que U peut prendre l'une des deux formes :

$$U_1 = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}, \quad U_2 = \begin{pmatrix} \alpha & \beta \\ \beta & -\alpha \end{pmatrix},$$

avec $\alpha^2 + \beta^2 \neq 0$.

Les matrices U_1 (resp. U_2) sont celles des similitudes directes (resp. inverses). On déduit de ces formules que le groupe $GO^+(\mathbf{R}^2)$ des similitudes directes est *commutatif*, donc aussi le groupe $O^+(\mathbf{R}^2)$ des rotations ; $GO^+(\mathbf{R}^2)$ opère de façon *simplement transitive* dans $\mathbf{R}^2 - \{0\}$. On voit aussi que :

$$GO^+(\mathbf{R}^2) \cup \{0\} \subset M_2(\mathbf{R})$$

est un *sous-corps* commutatif de l'anneau $M_2(\mathbf{R})$; il s'identifie au corps C des nombres complexes en identifiant la matrice :

$$\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

au nombre complexe $\alpha + \beta i$, image du vecteur de base e_1 par la similitude correspondante. Le groupe $O^+(\mathbf{R}^2)$ est alors identifié ainsi au groupe multiplicatif U des nombres complexes de module 1.

On appelle *groupe des angles* un groupe $\mathfrak{U}$ isomorphe à $O^+(\mathbf{R}^2)$ (donc à U) mais noté *additivement* ; il n'y a, par suite, pas de distinction essentielle à faire entre les notations d'angle et les notations de rotation plane, bien qu'il soit commode de parler de la « rotation d'angle θ » et de la noter :

$$\mathbf{r}(\theta) \in O^+(\mathbf{R}^2) = O^+(2, \mathbf{R}).$$

Puisque, par définition, r est un isomorphisme de $\mathfrak{U}$ sur $O^+(2, \mathbf{R})$, on a :

$$\mathbf{r}(\theta + \theta') = \mathbf{r}(\theta)\mathbf{r}(\theta'),$$
$$\mathbf{r}(0) = I \text{ (matrice unité)},$$
$$\mathbf{r}(-\theta) = (\mathbf{r}(\theta))^{-1}.$$

Par définition, les éléments α et β dans la matrice :

$$\mathbf{r}(\theta) = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

se notent $\cos\theta$ et $\sin\theta$ et s'appellent le cosinus et le sinus de l'angle $\theta \in \mathfrak{U}$. Les formules précédentes sur r se traduisent en les formules dites « trigonométriques » :

$$\cos(\theta + \theta') = \cos\theta \cos\theta' - \sin\theta \sin\theta',$$
$$\sin(\theta + \theta') = \sin\theta \cos\theta' + \cos\theta \sin\theta',$$
$$\cos 0 = 1,$$
$$\sin 0 = 0,$$
$$\cos(-\theta) = \cos\theta,$$
$$\sin(-\theta) = -\sin\theta,$$

qui ne font donc que transcrire des propriétés **du** groupe $O^+(2, R)$.

Pour deux vecteurs x et y de $\mathbf{R}^2$ de même longueur $\|x\| = \|y\| \neq 0$, il existe une rotation u et une seule telle que $u(x) = y$; l'angle θ de cette rotation est appelé l'angle *de y avec x* et noté $\widehat{(x, y)}$. Si les deux vecteurs sont unitaires, on a $\cos \theta = (x \mid y)$.

Si x, y, z sont trois vecteurs de même longueur dans $\mathbf{R}^2$, on a :

$$\widehat{(x, z)} = \widehat{(x, y)} + \widehat{(y, z)}.$$

Le groupe des angles II contient des éléments d'*ordre fini* : par exemple, l'angle *droit* δ qui correspond au nombre complexe $i \in U$ ou à la matrice :

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix};$$

on a $4\delta = 0$ (un angle de « quatre droits » est l'*angle nul*). Il n'est donc pas possible de définir sur II une relation d'ordre pour laquelle les relations $\theta > 0$, $\theta' > 0$ entraîneraient $\theta + \theta' > 0$, et il est absurde de parler d'un angle « plus petit qu'un autre ». Il est tout aussi absurde de considérer un angle comme une « grandeur mesurable », puisqu'on sait que, pour de telles grandeurs, il y a une relation d'ordre du type précédent. Par contre, une propriété fondamentale du groupe U est l'existence d'un *homomorphisme continu* φ, noté :

$$t \mapsto e^{it},$$

du groupe additif R sur U, qui est automatiquement dérivable et est le seul homomorphisme continu tel que $\varphi'(0) = i$. Il est périodique et sa plus petite période positive est 2π (ce qui *définit* le nombre π). Le cosinus et le sinus d'un *nombre réel t se définissent* alors par :

$$\cos t = \mathrm{Re}(e^{it}), \quad \sin t = \mathrm{Im}(e^{it})\,;$$

l'angle ρ tel que $r(\rho) = e^i$ est appelé *radian* et, si, pour un angle θ, on a $r(\theta) = e^{it}$, on dit (improprement) que t est une « mesure en radians » de θ (il y en a une infinité différant de multiples entiers de 2π ; cf. EXPONENTIELLE ET LOGARITHME). On a vu plus haut (*Générateurs du groupe orthogonal*) que toute rotation $r(\theta)$ est produit de deux symétries orthogonales s_1, s_2 autour de deux droites D_1, D_2 ; si $x_1 \in D_1$ et $x_2 \in D_2$ ont la même longueur et si $\widehat{(x_1, x_2)} = \omega$, on a $\theta = 2\omega$. Notons enfin que $O^+(2, R)$ est le groupe des commutateurs de $O(2, R)$.

Structure des transformations orthogonales

Pour toute transformation orthogonale $u \in O(E)$, il y a une décomposition de E en sous-espaces *deux à deux orthogonaux* V, W, $P_1, P_2, \ldots, P_r$ stables par u et tels que :

a) la restriction de u à V est l'identité ;

b) la restriction de u à W est la symétrie $x \mapsto -x$;

c) chacun des P_j est un plan (espace de dimension 2) et la restriction u_j de u à P_j est une rotation distincte de l'identité et de $x \mapsto -x$.

Si Ψ_j est une isométrie de P_j sur $\mathbf{R}^2$, il existe un angle θ_j distinct de 0 et de 2δ tel que $u_j = \Psi_j^{-1} r(\theta_j) \Psi_j$, et θ_j est déterminé « au signe près » indépendamment du choix de Ψ_j ; *les valeurs propres* de u sont 1 (de multiplicité dim V), -1 (de multiplicité dim W) et les $e^{\pm i\theta_j}$ (ces dernières peuvent être multiples si $\theta_j = \pm \theta_k$ pour $j \neq k$).

On a $\det(u) = (-1)^{\dim W}$; par suite, si $u \in O^+(E)$ et si dim E est *impair* (resp. $u \notin O^+(E)$ et dim E *pair*), W est nécessairement de dimension paire (resp. impaire) ; donc V ne peut être réduit à 0, en d'autres termes il existe au moins un vecteur $x \neq 0$ *invariant* par u.

Simplicité du groupe $O^{+}(3, \mathbf{R})$

Montrons que tout sous-groupe distingué N de $O^{+}(3, \mathbf{R})$ non réduit à l'identité est nécessairement égal à $O^{+}(3, \mathbf{R})$. Supposons donc qu'il existe $u \neq 1_E$ dans N, de sorte que (cf. *supra, Structure des transformations orthogonales)* il existe une droite D dont tous les points sont invariants par u, et la restriction de u au plan $P = D^{\perp}$ est une rotation d'angle $\theta \neq 0$ (déterminé « au signe près »). Distinguons trois cas :

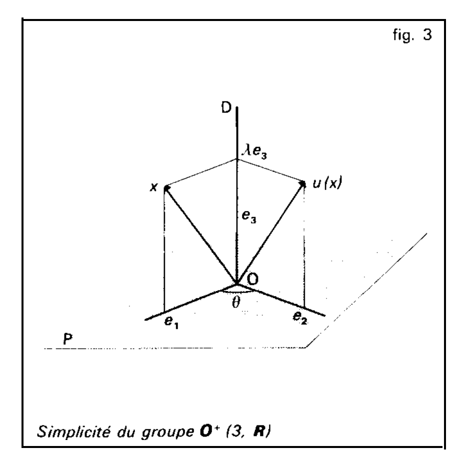
a) $\cos \theta = -1$ ou $\theta = 26$, autrement dit u est un *renversement* ; mais, comme N est distingué, il contient tous les renversements (cf. *supra, Propriétés de transitivité et de conjugaison,* in chap. 2), et donc il est égal *à* $O^{+}(3, \mathbf{R})$ (cf. *supra, Générateurs du groupe orthogonal).*

b) $\cos \theta < 0$. Soit e_3 un vecteur de longueur 1 dans D, e_1 un vecteur de longueur 1 dans P et $e_2 = u(e_1) \in P$; on a $(e_1 \ e_2) = \cos \theta < 0$. Considérons un vecteur $x = Ae, + e,$; on a $u(x) = \lambda e_3 + e_2$, donc $(x \mid u(x)) = \lambda^2 + \cos 8$, et, en prenant $A = (-\cos \theta)^{1/2}$, on obtient un vecteur tel que $(x \ u(x)) = 0$ (fig. 3). Soit alors v le renversement d'axe $\mathbf{R}x$; uvu^{-1} est le renversement d'axe Ru(x). Comme N est distingué,

$$v(uvu^{-1}) = (vuv^{-1})u^{-1} \in N,$$

et c'est le renversement d'axe orthogonal au plan $\mathbf{R}x \oplus \mathbf{R}u(x)$. On est ainsi ramené au cas *a*.

c) $0 < \cos \theta < 1$. On voit aisément qu'il existe un entier $n > 0$ tel que $\cos n\theta < 0$; comme $u^n \in N$, il suffit d'appliquer le cas *b* à u^n et la démonstration est achevée.



Simplicité du groupe $O^{+}(3, \mathbf{R})$

Les groupes $O^{+}(n, \mathbf{R})$ pour $n \geqslant 4$

En utilisant la simplicité du groupe $O^{+}(3, \mathbf{R})$, on peut, par un raisonnement tout aussi élémentaire mais assez long, prouver que :

$$O^{+}(n, \mathbf{R})/(Z_0 \cap O^{+}(n, \mathbf{R}))$$

est *simple pour* $n \geqslant 5$; cela entraîne que, si $n \geqslant 5$ est pair, il ne peut y avoir de sous-groupe Γ de $O^{+}(n, \mathbf{R})$ tel que $O^{+}(n, \mathbf{R})$ soit produit semi-direct de Z_0 et de Γ, car Γ serait d'indice 2, donc distingué. Par contre, le groupe $O^{+}(4, \mathbf{R})$ a une structure tout à fait exceptionnelle, liée à l'existence du *corps des quaternions* H (cf. ANNEAUX ET ALGÈBRES, chap. 2). Identifiant H et $\mathbf{R}^4$ on montre en effet que toute rotation de $\mathbf{R}^4$ peut s'écrire $x \mapsto sxt$, *où* s et t sont deux quaternions tels que $N(s)N(t) = 1$; en outre, si $sxt = s'xt'$ pour tout $x \in H$, on a nécessairement $s' = As$, $t' = A^{-1}t$ pour un $\lambda \in \mathbf{R}$. On en déduit que le groupe $O^{+}(4, \mathbf{R})/Z_0$ est isomorphe au produit de *deux groupes simples* isomorphes à $O^{+}(3, \mathbf{R})$; mais Z_0 n'est pas facteur direct dans $O^{+}(4, \mathbf{R})$.

Spineurs

L'algèbre des quaternions sur R se généralise de la façon suivante. Pour tout entier $n \geqslant 2$, il existe une algèbre C_n

sur R, de dimension 2^n, dite *algèbre de Clifford* d'indice n, qui est engendrée, en tant qu'algèbre. par l'élément unité 1 et n éléments e, ($1 \leqslant j \leqslant n$) identifiés à la base canonique de $\mathbf{R}^n$, et qui sont assujettis à vérifier les conditions suivantes :

$$e_i e_j = -e_j e_i, \quad \text{si } i \neq j,$$
$$e_i^2 = -1, \quad \text{pour tout } i.$$

On montre que les 2^n produits :

$$e_{i_1} e_{i_2} \dots e_{i_p}$$

(où $0 \leqslant p \leqslant n$, $i_1 < i_2 < \; < i_p$) forment une base sur R de l'espace vectoriel C,,. Ceux de ces éléments pour lesquels p est pair forment une sous-algèbre C_n^+de C_n, de rang $2^{n\ 1}$ sur R. Pour deux vecteurs a et x de $\mathbf{R}^n \subset C_n$, on a $ax + xa = -(x \mid a)$ dans C,,, donc $-2(a\ \ a) = a^2$ et finalement, si $a \neq 0$,

$$axa^{-1} = -x + 2\frac{(x|a)}{(a|a)}a\,;$$

ce qui prouve que l'application $x \mapsto -axa^{-1}$ de $\mathbf{R}^n$ dans lui-même n'est autre que la *réflexion orthogonale* s_a de droite Ra (cf. *supra, Générateurs du groupe orthogonal).*

Le *groupe multiplicatif* engendré dans C_n^+ par les produits ab, où a et b varient dans $R^n - \{0\}$ et sont de longueur 1, est noté Spin(n) ; on montre qu'il existe un homomorphisme surjectif et un seul σ : Spin(n) $\rightarrow$ O+(n, R) tel que $\sigma_{ab} = s_a s_b$; le noyau de cet homomorphisme est formé de l'identité et de -1, mais Spin(n) n'est pas produit semi-direct de ce sous-groupe et d'un groupe isomorphe à $O^+(n$, R). Lorsque l'on considère C_n^+ comme un espace vectoriel sur lequel Spin(n) opère par multiplication à gauche, les éléments de C_n^+ sont appelés *spineurs* (cf. la partie E ci-dessous Groupes de Lie).

3. Les groupes orthogonaux des formes non positives

Dans le chapitre 2, on peut remplacer, au départ, le produit scalaire par une forme bilinéaire symétrique non dégénérée *quelconque* $\Phi(x, y)$; pour une telle forme, il existe toujours au moins une base (dite *adaptée à* Φ) telle que :

$$\Phi(e_i, e_j) = 0 \quad \text{pour } i \neq j,$$
$$\Phi(e_i, e_i) = 1 \text{ pour} 1 \leqslant i \leqslant p,$$
$$\Phi(e_i, e_i) = -1 \text{ pour } p + 1 \leqslant i \leqslant n = \dim E,$$

et le nombre p est le même pour toutes les bases adaptées (« loi d'inertie ») ; on dit que $(p, n-p)$ est la *signature* de Φ ; un produit scalaire est donc une forme de signature $(n, 0)$.

La différence fondamentale entre le cas $1 < p < n$ et les cas $p = n$ et $p = 0$ réside dans l'existence de vecteurs $x \neq 0$ tels que $\Phi(x, x) = 0$, dits vecteurs *isotropes* (leur ensemble est appelé *cône isotrope* de E). Plus généralement, il y a des sous-espaces $V \neq \{0\}$ tels que la restriction de Φ *à* V soit *identiquement nulle* ; on dit que ces espaces sont *totalement isotropes* et leur dimension maximale est :

$$\nu = \sup(p, n-p),$$

appelée *indice de Witt* de Φ. On définit comme dans le chapitre 2 les notions de *vecteurs orthogonaux* (pour Φ) et de *sous-espaces orthogonaux* ; on a encore entre V et son orthogonal $V^\perp$ les mêmes relations, sauf les relations (équivalentes) $V \cap V^\perp = \{0\}$ et $V + V^\perp = E$; si $V \cap V^\perp \neq \{0\}$, on dit que V est un sous-espace *isotrope.* Dire que $V \subset V^\perp$ signifie que V est totalement isotrope ; pour tout sous-espace V, $V \cap V^\perp$ est totalement isotrope et c'est, en fait, le plus grand

sous-espace totalement isotrope contenu dans V ou $V^\perp$.

Ces notions d'« orthogonalité » relatives à Φ ont une traduction plus familière (tout au moins pour n = 4) en géométrie projective : si P(E) est l'espace projectif (de dimension n 1) associé à E, l'image Q dans P(E) du cône isotrope d'équation $\Phi(x, x) = 0$ est appelée *quadrique (ou hyperquudrique)* projective non dégénérée ; si x et y sont deux vecteurs $\neq 0$ dans E, orthogonaux pour Φ, on dit que les images de $\mathbf{R}x$ et Ry sont des points de P(E) conjugués *par rapport à Q*. Si D est une droite de E et H = $D^\perp$ l'hyperplan orthogonal pour Φ, on dit que le point de P(E) correspondant à D est le *pôle* de I'hyperplan projectif correspondant à H et que ce dernier est l'*hyperplan polaire* de ce point (par rapport à Q). Un sous-espace isotrope de E a pour image une variété projective *tangente à Q* et un sous-espace totalement isotrope a pour image une variété projective *contenue* dans Q (de dimension projective v 1 ; pour n = 4, $v = 2$, *ce* sont les *génératrice.~* de Q).

On définit les *similitudes* et les *transformutions orthogonales* relatives à Φ en remplaçant, dans les définitions du chapitre 2, le produit scalaire par $\Phi(x, y)$; on note GO(@) [resp. $O(\Phi)$] le groupe des similitudes [resp. le groupe orthogonal] relatif à Φ ; on a $GO(-\Phi) = GO(\Phi)$. La loi d'inertie montre que le multiplicateur $\mu(u)$ d'une similitude est nécessairement > 0 sauf si n est pair et $p = n/2$; sauf dans ce dernier cas, GO(@) est produit direct de $O(\Phi)$ et de $Z^+(E)$; si $p = n/2$, *ce* produit direct est un sous-groupe d'indice 2 (non facteur direct) dans GO(@).

Si on rapporte E à une base adéquate pour Φ, la matrice U d'une similitude relative à Φ est caractérisée par la relation :

$$ {}^tU . D . U = \mu D, $$

où :

$$ D = \text{diag}(\underbrace{1, \dots, 1}_{p}, \underbrace{-1, \dots, -1}_{n-p}) ; $$

on a donc $(\det U)^2 = \mu^n$, et en particulier, $\det U = \pm 1$ pour une transformation orthogonale ; on définit comme dans le chapitre 2 le groupe des rotations $O^+(\Phi)$; sauf lorsque n est pair et $p = n/2$, on dit que le sous-groupe d'indice 2 dans GO(@),

$$ GO^+(\Phi) = O^+(\Phi) \times Z^+(E), $$

est formé de similitudes directes (les autres étant dites inverses). Pour $n = 2p$, le groupe GO+(@) des similitudes directes est défini comme formé des similitudes u telles que $\det(u) = (\mu(u))^p$; il contient le produit direct précédent, qui en est un sous-groupe d'indice 2.

Le groupe $O(\Phi)$ relatif au cas $n = 4$, $p = 3$ joue un rôle fondamental en relativité restreinte et est connu sous le nom de *groupe de Lorentz* (le cône isotrope étant alors souvent appelé « cône de lumière »).

Nous supposons dans tout ce qui suit que $1 \leqslant p \leqslant n$ 1 ; on peut d'ailleurs supposer $p \geqslant n/2$ pour l'étude de GO(@).

Générateurs de $O(\Phi)$

Les *involutions* de $O(\Phi)$ se caractérisent comme ci-dessus (cf. *Générateurs du groupe orthogonal,* in chap. 2) ; mais, comme V^+ et V^- doivent être orthogonaux et tels que $V^+ + V^- = E$, ce sont nécessairement des espaces *non isotropes.* Si n = dim E, il est encore exact que toute transformation orthogonale est produit de n réflexions orthogonales au plus et que toute rotation est produit d'un nombre fini de renversements. Le centre Z_0 de $O(\Phi)$

est le même que dans le cas euclidien, et on a $Z_0 \subset O^+(\Phi)$ si n est pair, $O(\Phi) = Z_0 \times O^+(\Phi)$ si n est impair.

Propriétés de transitivité

Les propriétés de transitivité sont très différentes du cas euclidien. Pour un sous-espace V de E, de dimension m, on considère le sous-espace totalement isotrope $V \cap V^\perp$ de dimension $r \leqslant m$, puis un supplémentaire W de $V \cap V^\perp$ dans V ; la restriction de Φ à W est non dégénérée, soit $(q, m - r - q)$ sa signature. On a ainsi attaché trois invariants numériques m, r et q au sous-espace V ; étant donné deux sous-espaces V_1, V_2 de E, pour qu'il existe une transformation orthogonale u telle que $u(V_1) = V_2$, il faut et il suffit que ces trois entiers soient les mêmes pour V_1 et V_2. Lorsqu'il en est ainsi, il existe même une *rotation* u telle que $u(V_1) = V_2$, sauf dans un cas, celui où n est pair, Φ de signature $(n/2, n/2)$ et où il s'agit de sous-espaces *totalement isotropes* de dimension maximale $n/2$. En effet, il existe *deus* classes d'intransitivité $\mathfrak{n}_1$, $\mathfrak{n}_2$ de ces sous-espaces, pour l'action du groupe $O^+(\Phi)$; dans l'interprétation projective, ce sont les variétés projectives de dimension $n/2 - 1$ contenues dans la quadrique Q de dimension $n - 1$ et pour $n = 4$, on retrouve les deux systèmes de génératrices classiques. Si V, V' sont deux sous-espaces totalement isotropes de dimension $n/2$, dim $(V \cap V')$ a *même parité* que $n/2$ si V et V' appartiennent à la même classe d'intransitivité $\mathfrak{n}_i$ $(i = 1, 2)$, une *parité opposée à* celle de $n/2$ dans le cas contraire.

Le groupe $O(\Phi)$ pour $n = 2$

Le seul cas à considérer est celui de la signature (1, 1) : E, muni de Φ, est alors appelé *plan hyperbolique.* Il y a deux droites isotropes D_1, D_2 dans E ; une base (a_1, a_2) de E telle que $a_1 \in D_1$, $a_2 \in D_2$ et $\Phi(a_1, a_2) = 1$ est dite *base isotrope* de E. Par rapport à une telle base, la matrice d'une similitude a l'une des deux formes :

$$U_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}, \quad U_2 = \begin{pmatrix} 0 & \mu \\ \lambda & 0 \end{pmatrix}, \quad \lambda\mu \neq 0.$$

Les matrices U_1 (resp. U_2) sont celles des similitudes directes (resp. inverses) ; elles laissent invariantes chacune des droites isotropes (resp. les échangent). Le groupe $GO^+(\Phi)$ est donc encore *commutatif*, mais isomorphe au produit $R^* \times R^*$; il opère, dans ce cas, de façon simplement transitive dans $\mathbf{R}^2 - (D_1 \cup D_2)$, et le sous-anneau qu'il engendre dans $M_2(R)$ est isomorphe à $R \times R$. Le groupe $O^+(\Phi)$ est formé des matrices U_1 telles que $\lambda\mu = 1$ et est isomorphe au groupe multiplicatif R^* ; il contient donc un sous-groupe d'indice 2, $O^{++}(\Phi)$, dit groupe des rotations *orthochrones*, correspondant aux matrices pour lesquelles $\lambda > 0$ et isomorphe à $\mathbf{R}^*_+$. On a par suite un isomorphisme bicontinu :

$$t \mapsto \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$$

du groupe additif R sur le groupe $O^{++}(\Phi)$. Ce dernier opère de façon simplement transitive dans chacun des quatre « quadrants » ouverts déterminés dans $\mathbf{R}^2$ par les droites D_1, D_2. Si A et A' sont deux demi-droites contenues dans l'un d'eux (par exemple celui défini par $\xi_1 > 0$, $\xi_2 > 0$) la rotation orthochrone u telle que $u(A) = A'$ correspond à un nombre $\lambda > 0$ tel que $\lambda^{-2} = (D_1D_2DD')$, *birapport* de D_1, de D_2 et des droites D, D' contenant A, A' ; on est donc conduit ici à appeler « angle hyperbolique » (x, x') d'un vecteur

$x \in$ A et d'un vecteur $x' \in$ A' le nombre réel :

$$\frac{1}{2} \ln (D_1 D_2 D D') ;$$

on a encore la relation :

$$(\widehat{x, x''}) = (\widehat{x, x'}) + (\widehat{x', x''})$$

pour trois vecteurs du même quadrant. Si on rapporte $\mathbf{R}^2$ à une base adaptée, la matrice correspondant au nombre t s'écrit :

$$\begin{pmatrix} \text{ch}\, t & \text{sh}\, t \\ \text{sh}\, t & \text{ch}\, t \end{pmatrix}$$

et les traductions du fait que l'application de R sur $\mathbf{O}^{++}(\Phi)$ définie ainsi est un homomorphisme donnent cette fois les formules de la « trigonométrie hyperbolique ».

Les groupes $O^+(Q)$ pour $n \geqslant 3$

La description générale des rotations, en raison de l'existence des vecteurs isotropes, est ici beaucoup plus compliquée que celle donnée plus haut (cf. *Structure des transformations orthogonales*, in chap. 2) pour le cas euclidien, et nous ne l'indiquerons pas.

Le groupe des commutateurs de $\mathbf{O}(\Phi)$ est ici un sous-groupe d'indice 2 de $\mathbf{O}^+(\Phi)$, qu'on appelle encore le *groupe orthochrone* et qu'on note $\mathbf{O}^{++}(\Phi)$; si l'on écrit :

$$\begin{pmatrix} X & Y \\ Z & T \end{pmatrix}$$

une matrice de $\mathbf{O}^+(\Phi)$ par rapport à une base adaptée, X étant d'ordre p, le groupe $\mathbf{O}^{++}(\Phi)$ est formé des matrices pour lesquelles $\det(X) > 0$. On a $Z_0 \subset \mathbf{O}^{++}(\Phi)$ si p et $n-p$ sont pairs, $\mathbf{O}^+(\Phi) = Z_0 \times \mathbf{O}^{++}(\Phi)$ si n est pair et p impair.

On peut encore prouver que le groupe $\mathbf{O}^{++}(\Phi)/(Z_0 \cap \mathbf{O}^{++}(\Phi))$ est *simple* pour $n = 3$ et $n \geqslant 5$ ainsi que pour $n = 4$ et $p = 1$ ou $p = 3$; par contre, $\mathbf{O}^{++}(\Phi)/Z_0$ est produit direct de deux groupes simples pour $n = 4$ et $p = 2$. Les méthodes sont ici tout à fait différentes de celles qui sont employées dans le cas euclidien.

Géométries non euclidiennes

Soit Φ une forme bilinéaire symétrique de signature $(n-1, 1)$ sur E, et soit F la partie de l'espace projectif P(E) correspondant aux vecteurs x $\in$ E tels que $\Phi(x, x) < 0$; il résulte de la loi d'inertie que le groupe $\mathbf{O}(\Phi)$ opère transitivement sur F, et y définit donc une «géométrie » qu'on appelle *géométrie non euclidienne hyperbolique* (en dimension $n-1$) ; une *variété linéaire non euclidienne* de dimension $m \leqslant n-1$ est, par définition, l'intersection de F et d'une variété linéaire projective de dimension m qui *rencontre* F. Le groupe $\mathbf{O}(\Phi)$ opère alors encore *transitivement* sur les variétés linéaires non euclidiennes de dimension donnée m ; cela résulte des propriétés de transitivité (cf. *Propriétés de transitivité*, in chap. 3) et de l'hypothèse $p = n - 1$ (la propriété analogue serait inexacte pour $m \neq 0$ et pour $m \neq n - 1$ si l'on prenait $n/2 \leqslant p \leqslant n-2$). La quadrique Q est qualifiée d'« *absolu* » de l'espace non euclidien F ; elle n'y est évidemment pas contenue, mais donne un moyen commode d'étudier les propriétés de l'espace non euclidien F. Par exemple, comme deux plans de E peuvent rencontrer F et avoir une intersection qui soit une droite $\mathbf{R}x$ avec $\Phi(x, x) > 0$, l'existence d'une infinité de droites non euclidiennes passant par un point A, contenues dans le plan non euclidien déterminé par A et une droite D, et *ne rencontrant pas* D (dans F) est immédiate. De même,

l'étude faite plus haut (cf. *Le groupe* $O(\Phi)$ pour $n = 2$, in chap. 3) conduit à définir la *distance* non *euclidienne* de deux points A_1 et A_2 de F comme suit. *On* considère les deux points I et I′ où la droite non euclidienne A_1A_2 rencontre Q, et on prend pour distance de A_1 et de A_2 le nombre $|\ln(II'A_1A_2)|$, à un facteur près. Enfin, soit D_1 et D_2 deux droites non euclidiennes passant par un même point A de F ; si L est la droite de E correspondant à A, D_1 et D_2 correspondent à deux plans P_1 et P_2 de E contenant L ; par la loi d'inertie, la restriction de Φ à l'hyperplan H orthogonal à L est positive non dégénérée, autrement dit H est un espace euclidien de dimension $n - 1$; si x_1 et x_2 sont deux vecteurs de longueur 1 orthogonaux à L dans P_1 et P_2 respectivement, l'angle (x_1, x_2) a donc un sens, et c'est par définition l'*angle* (non euclidien) des « vecteurs » $\overrightarrow{AM_1}$ et $\overrightarrow{AM_2}$ si M_1 et M_2 correspondent à $\mathbf{R}x_1$ et à $\mathbf{R}x_2$ dans P(E).

L'espace non euclidien F défini ci-dessus est encore appelé *modèle de Cayley* de la géométrie non euclidienne hyperbolique. Comme $p = 1$, F est en correspondance biunivoque canonique avec la boule unité ouverte B de $\mathbf{R}^{n-1}$, à tout point $x \in B$ correspondant l'image dans F de la droite $\mathbf{R}(x + e_n)$ de $E = \mathbf{R}^n$; en transportant le modèle de Cayley par cette correspondance, on obtient le *modèle de Beltrami*, géométrie définie dans B, ou l'« absolu » est la sphère unité S définie par $\|x\| = 1$.

L'application :

$$x \mapsto \frac{2x}{1 + \|x\|^2}$$

est une *bijection* de la boule B sur elle-même, qui transforme les hyperplans de $\mathbf{R}^{n-1}$ rencontrant B en les sphères « orthogonales » à la sphère S. En transformant le modèle de Behrami par cette transformation, on obtient le *modèle de Klein-Poincaré* de la géométrie non euclidienne hyperbolique, où les « hyperplans non euclidiens » sont les traces sur B des sphères « orthogonales » à S (y compris les hyperplans diamétraux de S) ; l'intérêt de ce modèle est que l'angle de deux « droites non euclidiennes » (c'est-à-dire dans le modèle, deux cercles « orthogonaux » à S) est l'angle *euclidien* des tangentes à ces deux cercles en leur point commun. En transformant encore par une inversion de pôle situé sur S, on obtient comme modèle le *demi-espace de Poincaré*, ensemble des $x \in \mathbf{R}^{n-1}$ tels que $\xi_{n-1} > 0$, où les hyperplans non euclidiens sont les demi-sphères de centre dans l'hyperplan H défini par $\xi_{n-1} = 0$ et les hyperplans perpendiculaires à H.

Il y a une autre géométrie « non euclidienne » classique, la *géométrie elliptique de Riemann-Klein* ; ici, on prend pour Φ une forme bilinéaire symétrique de signature $(n, 0)$ (autrement dit, le produit scalaire euclidien dans E), et F = P(E), les variétés linéaires non euclidiennes étant ici simplement les variétés linéaires projectives ; il n'y a donc pas ici de « droites parallèles » non confondues et le groupe de la géométrie est $O(n, \mathbf{R})$. La notion d'angle (non euclidien) se définit comme dans la géométrie hyperbolique ; quant à la distance non euclidienne de deux points A_1 et A_2, on la définit comme la « mesure » en radians comprise entre 0 et $\pi/2$ de l'angle (euclidien) de deux vecteurs x_1 et x_2 correspondant respectivement à A_1 et à A_2.

4. Généralisations

Les groupes GL(E) et SL(E) se définissent de la même manière lorsque E est un

espace vectoriel de dimension finie sur un *corps commutatif* K quelconque ; si $n =$ dim E, on note aussi ces groupes $\mathbf{GL}(n, K)$ et $\mathbf{SL}(n, K)$. Tout ce qui a été vu dans le chapitre 1 pour le cas K = R s'étend sans changement au cas général, sauf en ce qui concerne la détermination des involutions lorsque K est de caractéristique 2 et, en ce qui concerne les propriétés de conjugaison, lorsque dim E = 2. On peut toutefois montrer que SL(E) est encore son propre groupe des commutateurs sauf lorsque dim E = 2 et que K est un corps fini ayant deux ou trois éléments ; la démonstration de simplicité faite plus haut dans le chapitre 1 s'applique alors sans modification et prouve que SL(E)/(Z $\cap$ SL(E)) est un groupe *simple,* sauf dans les deux cas précédents.

On peut étendre la définition de GL(E) et de SL(E) au cas où E est un espace vectoriel (à gauche) de dimension finie sur un corps K *non commutatif,* mais il faut utiliser dans ce cas une autre définition du déterminant ; moyennant quoi, on peut encore prouver la simplicité du groupe $\mathbf{SL}$(E)/(Z $\cap$ SL(E)) ; le centre Z(E) de GL(E) est ici formé des homothéties $x \mapsto \lambda x$ où $\lambda \neq 0$ est dans le *centre* de K.

La définition de GL(E) est aussi valable pour un module (à gauche) E sur un anneau quelconque A ; mais, ici, la structure de ce groupe dépend de façon essentielle de la structure de l'anneau A, et on ne connaît de résultats satisfaisants que dans un petit nombre de cas particuliers.

La notion de groupe orthogonal $\mathbf{O}(\Phi)$ se généralise aussi au cas où E est un espace vectoriel de dimension finie sur un corps commutatif K, que nous supposerons en outre *de caractéristique* $\neq 2$ (la caractéristique 2 introduit ici des phénomènes spéciaux) ; Φ est une forme bilinéaire symétrique non dégénérée sur E. Il faut remarquer d'abord qu'il y a toujours ici des bases *orthogonales* (e_i) de E pour la forme Φ, c'est-à-dire $\Phi(e_i, e_j) = 0$ pour $i \neq j$; mais, si l'on pose $\Phi(e_i, e_i) = \alpha_i$, il n'est pas possible en général d'obtenir une base orthogonale pour laquelle $\alpha_i = \pm 1$ pour tout i ; la notion de signature de Φ n'a pas de sens lorsque K n'est pas un corps ordonné. Par contre, la définition des vecteurs et sous-espaces isotropes subsiste sans modification ; on appelle encore *indice de Witt* de Φ la dimension maximale ν des sous-espaces totalement isotropes, et on a $2\nu \leqslant n$. Il faut noter que, lorsque K est algébriquement clos (par exemple K = C), on a toujours $\nu = [n/2]$, partie entière de $n/2$; si K est fini, on a $\nu = [n/2]$ pour n impair, $\nu = n/2$ ou $n/2 - 1$ si n est pair.

Tout ce qui a été dit dans le chapitre 3 sur les involutions de $\mathbf{O}(\Phi)$ subsiste sans changement dans le cas général. Les questions de transitivité sont résolues par le *théorème de Witt* : Soit deux sous-espaces vectoriels V_1 et V_2 de même dimension dans E, pour qu'il existe une transformation $u \in \mathbf{O}(\Phi)$ telle que $u(V_1) = V_2$, il faut et il suffit que les restrictions de Φ à V_1 et à V_2 soient des formes équivalentes (dégénérées ou non). On peut encore alors transformer V_1 en V_2 par une rotation, sauf dans le même cas d'exception que pour K = R (cf. *Propriétés de transitivité,* in chap. 3).

Le groupe $\mathbf{O}^+(\Phi)$ est encore commutatif pour $n = 2$; il est formé des matrices :

$$\begin{pmatrix} \xi & -\frac{\alpha_2}{\alpha_1}\eta \\ \eta & \xi \end{pmatrix}$$

telles que $\alpha_1\xi^2 + \alpha_2\eta^2 = \alpha_1$; si $-\alpha_2/\alpha_1$ n'est pas un carré dans K, ce groupe est isomorphe au groupe multiplicatif des éléments de norme 1 dans l'extension

quadratique $K(\sqrt{-\alpha_2/\alpha_1})$ de K ; sinon, il est isomorphe à $K^* \times K^*$.

Il n'y a rien d'analogue en général à la prétendue « mesure » des angles ; autrement dit, il n'existe pas en général d'homomorphisme du groupe additif de K sur le groupe des rotations $\mathbf{O}^+(\Phi)$.

Lorsque $n \geqslant 3$, il faut distinguer, dans l'étude de la structure du groupe $\mathbf{O}^+(\Phi)$, le cas $\nu \geqslant 1$ et le cas $\nu = 0$.

Pour $\nu \geqslant 1$ (autrement dit, lorsqu'il existe des vecteurs isotropes $\neq 0$), on considère le groupe $\Omega(\Phi) \subset \mathbf{O}^+(\Phi)$ des commutateurs de $\mathbf{O}(\Phi)$. Le quotient $\mathbf{O}^+(\Phi)/\Omega(\Phi)$ est isomorphe à K^*/K^{*2}, en désignant par K^{*2} le groupe des carrés des éléments $\neq 0$ de K ; si n est pair, le groupe $\Omega(\Phi) \cap Z_0$ est égal à Z_0 si et seulement si le discriminant de Φ est un carré dans K. Lc groupe :

$$\Omega(\Phi)/(\Omega(\Phi) \cap Z_0)$$

est *simple* pour $n \geqslant 5$. Pour $n = 3$, $\Omega(\Phi)$ est isomorphe à $\mathbf{SL}(E)/Z_0$, donc simple sauf si K a trois éléments ; pour $n = 4$ et $\nu = 1$, le discriminant Δ de Φ ne peut être un carré dans K ; $\Omega(\Phi)$ est alors simple et isomorphe à SL(F), où F est un espace vectoriel de dimension 2 sur $K(\sqrt{\Delta})$. Enfin, pour $n = 4$ et $\nu = 2$, le groupe :

$$\Omega(\Phi)/(\Omega(\Phi) \cap Z_0)$$

est isomorphe au produit :

$$(\mathbf{SL}(E)/Z_0) \times (\mathbf{SL}(E)/Z_0),$$

dont les facteurs sont simples si $K \neq F_3$.

Le cas $\nu = 0$ (absence de vecteurs isotropes $\neq 0$) est tout à fait différent, et la structure de $\mathbf{O}(\Phi)$ *dépend essentiellement du corps de base* K. Prenons par exemple $n = 3$ et. pour Φ. le produit scalaire usuel :

$$\Phi(x, y) = \xi_1\eta_1 + \xi_2\eta_2 + \xi_3\eta_3,$$

mais *sur le corps Q des nombres rationnels* ; les matrices $U = (a_{ij})$ ont donc leurs éléments *rationnels*, vérifiant en particulier les équations :

$$\alpha_i^1 + \alpha_i^2 + \alpha_i^3 = 1,$$

pour $i = 1, 2, 3$. Or, pour toute solution de l'équation :

$$r_1^2 + r_2^2 + r_3^2 = 1$$

en nombres rationnels (supposés réduits en fractions irréductibles), le dénominateur de chaque r_i *n'est pas divisible par 2.* Supposons en effet le contraire : on voit alors aussitôt qu'on aurait une relation de la forme :

$$m^2 2^{2k} = p_1^2 + p_2^2 + p_3^2,$$

ou k est un entier > 0, m un entier impair, p_1, p_2, p_3 des entiers dont l'un au moins est impair ; mais il est immédiat de vérifier que, dans ces conditions, la somme $p_1^2 + p_2^2 + p_3^2$ n'est jamais multiple de 4, d'où notre assertion.

Cela étant, pour tout entier $r \geqslant 1$, soit G_r le sous-groupe de $\mathbf{O}(\Phi)$ formé des matrices de la forme $I + 2^r V$, où V est une matrice à coefficients rationnels dont les dénominateurs ne sont pas divisibles par 2. Pour toute matrice $U \in \mathbf{O}(\Phi)$, on a alors :

$$U(I + 2^r V)U^{-1} \in G_r,$$

en vertu de ce qui précède. Par suite, chaque G_r est un sous-groupe *distingué* de $\mathbf{O}(\Phi)$; on montre qu'ils sont tous différents et forment une suite descendante :

$$\mathbf{O}(\Phi) \supset G_1 \supset G_2 \supset \dots \supset G_r \supset \dots$$

Si $K = \mathbf{Q}$, le phénomène précédent ne peut se produire que pour les dimensions 3 et 4 ; mais on peut, pour toute valeur de

II, donner des exemples de corps K et de forme Φ pour lequel on a une suite infinie décroissante de sous-groupes distingués de $O(\Phi)$.

5. Groupes symplectiques et groupes unitaires

Deux autres types de groupes « classiques » ont été étudiés depuis le milieu du XIXe siècle. Si E est un espace vectoriel sur un corps (commutatif) K de dimension finie n, une forme bilinéaire *alternée* Φ sur E ne peut être non dégénérée que si $n = 2\nu$ est *pair.* Il existe alors une base :

$$(e_i),\ 1 \leqslant i \leqslant 2\nu,$$

de E (dite base *symplectique)* telle que :

$$\Phi(e_i, e_{i+\nu}) = 1,$$

pour $1 \leqslant i \leqslant \nu$, et :

$$\Phi(e_i, e_j) = 0,$$

pour tout autre couple d'indices, de sorte que l'on a :

$$\Phi(x, y) = \sum_{i=1} (\xi_i \eta_{i+\nu} - \xi_{i+\nu} \eta_i)\ ;$$

toutes les formes alternées non dégénérées sont équivalentes. On appelle *groupe symplectique* (sur K) et on note $\mathbf{Sp}(2\nu, K)$ le sous-groupe formé des $u \in GL(E)$ tels que :

$$\Phi(u(x), u(y)) = \Phi(x, y),$$

pour x, y dans E.

Considérons maintenant un corps K (non nécessairement commutatif) muni d'une *involution* :

$$\xi \mapsto \xi^*,$$

distincte de l'identité, c'est-à-dire une bijection de K sur lui-même telle que :

$$\xi^{**} = \xi, \quad (\xi + \eta)^* = \xi^* + \eta^*,$$
$$(\xi\eta)^* = \eta^*\xi^*, \quad 1^* = 1\ ;$$

soit E un espace vectoriel à gauche de dimension finie sur K, et soit Φ une forme *hermitienne* non dégénérée sur E (pour l'involution donnée) ; le groupe $U(\Phi)$ des $u \in GL(E)$ tels que :

$$\Phi(u(x), u(y)) = \Phi(x, y),$$

pour x, y dans E est appelé le *groupe unitaire* relatif à Φ. Lorsque K est commutatif, les $u \in U(\Phi)$ tels que $(\det u)(\det u)^* = 1$ forment un sous-groupe distingué $U^+(\Phi)$, ou SU(@), dit *groupe spécial unitaire.*

Pour le groupe symplectique et pour le groupe unitaire, les définitions des vecteurs et des sous-espaces orthogonaux, des sous-espaces totalement isotropes, des sous-espaces isotropes et de l'indice de Witt sont les mêmes que dans le chapitre 4. Pour le groupe symplectique, tout vecteur est isotrope et l'indice de Witt est $n/2$; les transvections (cf. *Générateurs,* in chap. 1) appartenant à $\mathbf{Sp}(2\nu, K)$ sont celles pour lesquelles l'hyperplan H de la transvection (toujours isotrope) est orthogonal à la droite D de la transvection ; les transvections symplectiques *engendrent* $\mathbf{Sp}(2\nu, K)$, donc $\det(u) = 1$ pour tout $u \in \mathbf{Sp}(2\nu, K)$. Le centre de $\mathbf{Sp}(2\nu, K)$ est Z_0. Pour p premier, désignons par F_p le corps fini des entiers relatifs modulo p. Le groupe $\mathbf{Sp}(2\nu, K)/Z_0$ est *simple* pour $\nu \geqslant 1$, sauf lorsque $\nu = 1$ et $K = \mathbf{F}_2$ ou $K = \mathbf{F}_3$, et lorsque $\nu = 2$ et $K = \mathbf{F}_2$; dans ce dernier cas, $\mathbf{Sp}(4, \mathbf{F}_2)/Z_0$ est isomorphe au groupe symétrique Σ_6. Pour $\nu = 1$, $\mathbf{Sp}(2, K) = SL(2, K)$.

Lorsque $n = 4$, les plans totalement isotropes de E correspondent, dans

l'espace projectif P(E), aux droites d'un *complexe linéaire,* et le groupe $Sp(4, K)/Z_0$ est le groupe qui laisse ce complexe invariant, d'où son nom.

Pour les groupes unitaires, on a $v \leqslant n/2$, mais v peut prendre toutes les valeurs entières remplissant cette condition. Pour $v > 0$, il existe dans $U(\Phi)$ des transvections dont l'hyperplan H est nécessairement isotrope et la droite D est orthogonale à H. Ces transvections engendrent un sous-groupe distingué $T(\Phi)$ de $U(\Phi)$ dont le centre $W(\Phi)$ est égal à $T(\Phi) \cap Z(E)$; le groupe $T(\Phi)/W(\Phi)$ est *simple,* sauf lorsque $K = \mathbf{F}_9$ et $n = 2$, ou lorsque $K = \mathbf{F}_4$, $n = 2$ et $n = 3$; $T(\Phi)$, pour $n = 2$ et K commutatif, s'identifie à SL(2, K,), où K_0 est le sous-corps des invariants de l'involution de K. Lorsque K est commutatif, on a $T(\Phi) = U^+(\Phi)$, sauf lorsque $n = 3$ et $K = \mathbf{F}_4$. Lorsque $v \geqslant 2$, ou lorsque $n \geqslant 3$ et que K est de rang fini sur son centre, $T(\Phi)$ est le groupe des commutateurs de $U(\Phi)$; par contre, pour $n = 2$, il y a des corps non commutatifs de rang 4 sur leur centre tels que $T(\Phi)$ ne soit pas le groupe des commutateurs de $U(\Phi)$, le quotient $U(\Phi)/T(\Phi)$ pouvant avoir des facteurs de composition simples (non commutatifs). Pour $v = 0$, les mêmes phénomènes que pour les groupes orthogonaux peuvent se produire.

JEAN DIEUDONNÉ

Bibliographie

E. ARTIN, *Geometric Algebra*, Wiley. New York, 1988 / R. DEHEUVELS, *Formes quadratiques et groupes classiques*, P.U.F., 1981 / J. DIEUDONNÉ, *Sui les groupes classiques*, Hermann, Paris, réimpr. 198 1 ; *Algèbre linéaire et géométrie élémentaire*, 3' éd., Hermann, Paris, 1968.

C. Groupes finis

Née de l'étude des groupes de permutations des racines d'équations, la théorie des groupes finis s'est développée indépendamment depuis le *Traité des substitutions et des équations algébriques* (1870) de Camille Jordan. Après les travaux importants de Burnside, de Frobenius et de leurs élèves vers le commencement du XXe siècle, cette théorie connut une période de développement lent, faute de méthodes pour résoudre les nombreux problèmes posés par ces pionniers. Les efforts de mathématiciens comme P. Hall et R. Brauer pendant cette période ont engendré les nouvelles méthodes qui, après 1955, ont amené une intense activité dans ce domaine ; des progrès énormes ont été accomplis, particulièrement dans la théorie des groupes simples et la théorie des relations entre un groupe et ses sous-groupes. Mais beaucoup de questions sont restées longtemps ouvertes et sont l'objet d'une recherche acharnée.

1. Groupes de permutations

Historiquement la théorie des groupes finis commença avec l'étude des groupes symétriques et de leurs sous-groupes, les groupes de permutations. Soit E un ensemble fini formé des n éléments $e_1, \dots, e_n$, $n \geqslant 1$. Une *permutation* π *des éléments* $e_1, \dots, e_n$ *(ou* encore une *permutation* π *sur* E) est une application $x \mapsto \pi(x)$ de E dans E, telle que chaque élément y de E soit l'image $y = \pi(x)$ d'un élément *unique* x de E. L'application π^{-1}, envoyant chaque élément y sur l'élément x tel que $y = \pi(x)$, est alors aussi une permutation sur E, qui s'appelle *l'inverse* de π. Le *produit* $\pi\rho$ de deux permutations π, ρ sur E est la

permutation de $e_1, \ldots, e_n$ définie par : $\pi\rho(x) = \pi(\rho(x))$, pour **tout** x dans E. Avec ces définitions de l'inversion et de la multiplication, l'ensemble des permutations sur E forme un groupe fini C(E), le groupe *symétrique* de E. Son élément neutre est la permutation *identité* $1 = 1_E$ sur E, qui envoie chaque $x = e_1, \ldots, e_n$ sur lui-même : $1_E(x) = x$.

Le groupe symétrique C(E) est déterminé à un isomorphisme près par le nombre $n = |E|$ d'éléments de E ; et on l'appelle souvent le *groupe symétrique* Σ_n *de degré* n, sans spécifier l'ensemble E. L'ordre $|\Sigma_n|$ du groupe Σ_n, c'est-à-dire le nombre de ses éléments, est $n! = 1.2 \ldots (n-1).n$.

On peut représenter une permutation π de $e_1, \ldots, e_n$ graphiquement par un tableau formé des éléments $e_1, \ldots, e_n$ et de flèches. Chaque flèche joint un élément $x = e_1, \ldots, e_n$ à son image $y = \pi(x)$. Par exemple, si π est la permutation des éléments a, b, c, d, définie par $\pi(a) = c$, $\pi(b) = a$, $\pi(c) = d$, $\pi(d) = b$, son tableau est :

$$\begin{array}{ccc} a & \rightarrow & c \\ \uparrow & & \downarrow \\ b & \leftarrow & d \end{array}$$

Une permutation comme celle-ci, dont le tableau a la forme d'une seule boucle, s'appelle une *permutation cyclique*. Elle se note en donnant les éléments dans leur ordre cyclique $\pi = (a, c, d, b)$. Dans cette écriture, on peut commencer avec n'importe quel élément et écrire π sous les formes équivalentes :

$$\pi = (a, c, d, b) = (c, d, b, a) = (d, b, a, c) = (b, a, c, d).$$

Toute permutation sur E s'écrit comme un produit de permutations cycliques sur certains sous-ensembles de E. Ainsi, la permutation π de $1, 2, \ldots, 6$ définie par : $\pi(1) = 5$, $\pi(2) = 6$, $\pi(3) = 3$, $\pi(4) = 2$, $\pi(5) = 1$, $\pi(6) = 4$, a pour tableau :

$$1 \leftrightarrows 5 \qquad \begin{array}{ccc} 2 & \rightarrow & 6 \\ & \nwarrow \swarrow & \\ & 4 & \end{array} \qquad 3 \circlearrowleft$$

Il est évident que l'ensemble $\{1, 2, \ldots, 6\}$ se décompose en une réunion disjointe des sous-ensembles $\{1, 5\}$, $\{2, 6, 4\}$ et $\{3\}$, sur lesquels π opère comme les permutations cycliques (1, 5) (2, 6, 4) et (3). Ces permutations cycliques sont les *cycles* de π. La *longueur* d'un cycle de π est le nombre d'éléments dans le sous-ensemble correspondant. Donc π a des cycles de longueur 2, 3 et 1. On écrit π comme le produit (dans n'importe quel ordre) de ses cycles : $\pi = (1, 5)(2, 6, 4)(3) = (4, 2, 6)(3)(1, 5)$.

On a l'habitude de supprimer les cycles de longueur 1 quand l'ensemble E est connu. On écrit ainsi (1, 5)(2, 6, 4) au lieu de (1, 5)(2, 6, 4)(3) pour la permutation considérée précédemment. Cette notation est cohérente avec la notation adoptée pour la multiplication, car la permutation $(1, 5)(2, 6, 4) = (1, 5)(2, 6, 4)(3)$ est en fait le produit des permutations $(1, 5) = (1, 5)(2)(3)(4)(6)$ et $(2, 6, 4) = (2, 6, 4)(1)(3)(5)$.

Si les longueurs des cycles d'une permutation π sont $l_1, \ldots, l_k$ alors la *signature* $\mathrm{sgn}(\pi)$ de π est le nombre :

(1) $$\mathrm{sgn}(\pi) = (-1)^{(l_1 + \cdots + l_k) + k},$$

donc la signature de (1, 5)(2, 6, 4) est $(-1)^{(2+3)+2} = -1$.

Si cette signature est égale à 1, la permutation π est *paire*, si elle est égale à -1, elle est *impaire*. La fonction sgn est une fonction *multiplicative* : $\mathrm{sgn}(\pi\rho) = \mathrm{sgn}(\pi)\mathrm{sgn}(\rho)$, pour toutes les permutations π et ρ. L'application $\pi \mapsto \mathrm{sgn}(\pi)$ est

donc un morphisme surjectif du groupe Σ_n sur le groupe multiplicatif des nombres ± 1 (si $n \geqslant 2$). Le noyau **de** ce morphisme, c'est-à-dire l'ensemble A,,, des permutations paires de Σ_n est alors un sous-groupe distingué de Σ_n. C'est le *groupe alterné de degré n.* Son ordre est A,, $| = n!/2$.

Les groupes Σ_n et A, sont des exemples de *groupes de permutations.* Un tel groupe G sur un ensemble E est un sous-groupe quelconque du groupe symétrique C(E). Le *degré,* deg G, de G est alors le nombre E d'éléments dans E. Son *ordre* G est le nombre de ses éléments.

À chaque élément x de E on associe son *stabilisateur* G_x, le sous-groupe de toutes les permutations π dans G envoyant x sur lui-même : $\pi(x) = x$.

Si, pour tout couple (x, y) d'éléments distincts de E, il existe au moins une permutation π dans G telle que $\pi(x)$ = y, on dit que le groupe *G* est *transit.$* Par exemple, le groupe alterné A,, est transitif pour $n \geqslant 3$: si x et y sont deux éléments distincts de E, il existe un élément z de E, distinct de x et de y. La permutation (x, y, z) est paire, d'après (1), et envoie x sur y.

On a déjà décomposé une permutation π sur un ensemble E en permutations cycliques sur certains sous-ensembles disjoints de E. De manière analogue, on peut décomposer un groupe G de permutations sur E en groupes transitifs de permutations sur certains sous-ensembles disjoints de E. À tout élément x de E, on associe sa *G-orbite* G(s), qui est l'ensemble de toutes les images $\pi(x)$ de x par les permutations π de G. L'élément $x =$ l(s) appartient à son orbite $G(x)$, et, pour tout élément $y \in$ G(x), on a $G(y)$ = G(x). L'ensemble E est donc la réunion disjointe des G-orbites E,, ..., E_k de ses éléments. Si une permutation π appartient à G, sa restriction π_i est une permutation sur l'orbite E,, pour i = 1, ... k. Ces restrictions π_1, ... π_k déterminent la permutation π. L'ensemble G, des restrictions π_i (à E,) des permutations π de G est un groupe transitif de permutations sur E_i, $i = 1, ..., k$; et l'application $\pi \mapsto \pi_i$ est un morphisme surjectif du groupe G sur G,. On a donc analysé le groupe G en les groupes transitifs G,, ... G_k au moyen des morphismes $\pi \mapsto \pi_i$.

La famille des groupes transitifs de permutations est *universelle,* en ce sens que chaque groupe fini H est isomorphe à un groupe G de cette famille. En effet, chaque élément σ de H détermine une permutation π_σ des éléments de H, définie par $\pi_\sigma(\tau) = \sigma\tau$, pour tout τ de H. L'application o- π_σ est un isomorphisme du groupe H sur un groupe transitif G de permutations sur H. Le groupe G ainsi obtenu a la propriété suivante : Si une permutation π de G laisse invariant au moins un élément de l'ensemble sur lequel G opère, alors π = 1. On appelle *régulier* tout groupe transitif de permutations ayant cette propriété. Tout groupe régulier G de permutations s'obtient à partir d'un groupe abstrait H de la manière décrite plus haut.

Comme la famille des groupes transitifs de permutations est universelle, il ne peut être question de les classer tous. On peut pourtant essayer de classer certaines sous-familles importantes de ces groupes. Une de ces sous-familles est celle des *groupes de Frohenius.* Un tel groupe G est un groupe transitif, mais non régulier, de permutations sur un ensemble E, avec la propriété suivante : Si une permutation π dans G laisse invariants au moins deux éléments de E, alors $\pi = 1$. Frobenius a montré (en 1901) qu'un groupe de Frobenius a toujours un unique sous-groupe régulier distingué K, appelé *noyau de Frohenius* de G, tel que $G = G_xK$ et $G_x \cap K = \{1\}$ pour

tout stabilisateur G_x d'un élément .Y dans E. En 1936, Zassenhaus a donné une classification complète des stabilisateurs G_x des groupes de Frobenius. Il n'y a pas de classification complète des noyaux K de Frobenius, mais Thompson (1959) a démontré une conjecture de Frobenius : tous ces noyaux sont des groupes nilpotents.

Une autre sous-famille est la famille des *groupes de Zassenhaus.* Un groupe de Zassenhaus G est un groupe transitif de permutations sur un ensemble E, tel que la restriction d'un stabilisateur G, soit un groupe de Frobenius sur l'ensemble formé de E moins l'élément x. Pour éviter les cas triviaux, on suppose aussi qu'il n'y a pas de sous-groupe régulier distingué dans G. Zassenhaus, Feit, Ito et Suzuki sont arrivés à une classification complète des groupes de Zassenhaus. Un tel groupe G est soit un PSL(2, K) pour un corps fini K (cf. *infra*, chap. 2), ou une extension de ce groupe par un groupe d'ordre 2, soit un des *groupes de Suzuki* (une autre famille de groupes simples découverte par Suzuki en 1960 lors de l'étude de ce problème).

Un groupe G de permutations sur un ensemble E est *n-fois transitif*, pour un entier positif n, si E $\geqslant n$, et si, chaque fois que l'on considère deux n-chaînes $x_1, \ldots, x_n$ et y,, ... y_n d'éléments de E où les $x_1, \ldots, x_n$ (resp. $y_1, \ldots, y_n$) sont tous distincts, il existe au moins une permutation π dans G, telle que $\pi(x_1) = y_1, \ldots, \pi(x_n) = y_n$. Les groupes de Zassenhaus sont des groupes *2-fois transitifs*. Le groupe symétrique Σ_n est n-fois transitif, pour tout $n \geqslant 1$. Le groupe alterné A,, est (n 2)-fois transitif pour $n \geqslant 3$. Il y a beaucoup de groupes 3-fois transitifs, les groupes PGL(2, K). par exemple, où K est un corps fini. Mais on ne connaît que quatre groupes 4-fois transitifs, autres que les groupes Σ_n pour $n \geqslant 4$, et A, pour $n \geqslant 6$. Ce sont les *groupes de Mathieu (1861* et 1873) $M_{24}, M_{23}, M_{12}, M_{11}$, dont les degrés et les ordres sont : deg (M_{24}) = 24, avec M_{24} = 244 823 040 ; deg (M_{23}) = 23. avec $|M_{23}|$ = 10200960; deg (M_{12}) = 12, avec $|M_{12}|$ = 95040; deg (M,,) = 11, avec M_{11} = 7 920. Les groupes M_{24} et M_{12} sont même 5-fois transitifs, mais non O-fois transitifs. On a de bonnes raisons de croire que tout groupe 6-fois transitif (ou plus) est Σ_n ou A,,, mais il n'y a aucune démonstration de cette conjecture.

2. Groupes simples

Si H est un sous-groupe distingué d'un groupe fini G, le morphisme surjectif naturel de G sur le groupe quotient G/H, ayant H pour noyau, nous donne une sorte d'analyse du groupe G en les deux groupes H et G/H. Les deux cas $H = \{1\}$, et $H =$ G sont triviaux, le groupe G étant alors isomorphe à l'un des deux groupes H et G/H. Dans tous les autres cas, les ordres G/H et H sont strictement plus petits que l'ordre de G. Les groupes G/H et H sont donc plus simples que G. Le groupe G est appelé *simple* si G $\neq \{1\}$ et si l'on ne peut pas l'analyser ainsi en des groupes d'ordre strictement plus petit, c'est-à-dire si $\{1\}$ et G sont les seuls sous-groupes distingués de G. Par exemple, pour chaque entier premier p, le groupe cyclique C_p d'ordre p est simple.

Tout groupe fini G peut se décomposer en groupes simples : si G = $\{1\}$, il n'y a rien à faire ; si G $\neq \{1\}$, il y a toujours un sous-groupe distingué H, de G tel que G/H_1 soit un groupe simple. Si H, = $\{1\}$, l'analyse est terminée. Sinon, il existe un sous-groupe distingué H_2 de H,, tel que H_1/H_2 soit un groupe simple. Si on itère

cette construction, on aboutit à une suite $G = H_0, H_1, H_2, ..., H_n = \{1\}$ de sous-groupes de G, où H_i est distingué dans H_{i-1}, et où H_{i-1}/H_i est un groupe simple, $i = 1, ..., n$. Une telle suite s'appelle une *suite de Jordan-Hölder* du groupe fini G, et les groupes quotients $H_0/H_1, H_1/H_2, ..., H_{n-1}/H_n$ s'appellent les *facteurs de Jordan-Hölder* de G. La terminologie adoptée ici suit N. Bourbaki ; les théoriciens des groupes finis, traditionnellement, continuent à réserver le terme de suite de composition à ce que nous appelons ici suite de Jordan-Holder.

Considérons, par exemple, le groupe symétrique Σ_4 des permutations de 1, 2, 3, 4. L'ordre de Σ_4 est 4 ! = 24. Le groupe alterné A_4, d'ordre 4 !/2 = 12 est un sous-groupe distingué dans Σ_4, et le groupe quotient Σ_4/A_4 est isomorphe au groupe simple C_2. Les trois permutations (12)(34), (13)(24), (14)(23) forment avec l'identité un sous-groupe V d'ordre 4 dans A_4 qui s'appelle le 4-groupe *de Klein*. Ce groupe V est distingué dans Σ_4, et donc dans A_4. Le groupe quotient A_4/V est isomorphe au groupe C_3. Le 4-groupe V est commutatif, et tous ses sous-groupes sont distingués. Pour $\sigma = (12)(34)$, ou $\sigma = (13)(24)$, ou $\sigma = (14)(23)$, le sous-groupe $\{\sigma, 1\}$ est isomorphe à C_2, ainsi que le groupe quotient $V/\{\sigma, 1\}$. On a donc construit pour Σ_4 une suite de Jordan-Holder Σ_4, A_4, V, $\{\sigma, 1\}$, $\{1\}$, ayant comme facteurs de Jordan-Holder C_2, C_3, C_2, C_2 (à des isomorphismes près).

Un groupe fini dont chaque facteur de Jordan-Holder est isomorphe à un C_p, où p est un nombre premier, est dit *résoluble*. Le groupe Σ_4 est donc résoluble.

Dans l'exemple ci-dessus, il y avait trois choix possibles pour le groupe $\{\sigma, 1\}$. Un groupe G peut donc avoir plusieurs suites de Jordan-Holder. Il y a malgré **tout** une certaine unicité des suites de Jordan-Hölder : *les facteurs de Jordan-Hölder de* G *sont indépendants du choix de la suite de Jordan-Hölder* (*théorème de Jordan-Hölder*), c'est-à-dire que, si $G = H_0, H_1, ..., H_n = \{1\}$, et $G = K_0, K_1, ..., K_m = \{1\}$ sont **deux** suites de Jordan-Holder de G, on a $n = m$, et il existe une permutation π de $\{1, ..., m = n\}$ telle que le groupe H_{i-1}/H_i soit isomorphe au groupe $K_{\pi(i)-1}/K_{\pi(i)}$, pour tout $i = 1, ..., m = n$. Chaque groupe fini peut donc être analysé en groupes simples uniques, qui sont ses facteurs de Jordan-Holder. D'où l'importance de l'étude des groupes simples.

Les premiers groupes simples non cycliques furent découverts dans la première moitié du XIXe siècle. Les groupes alternés A_n, sont des groupes simples, pour tout $n \geqslant 5$. C'est sur cette découverte que repose la démonstration moderne du théorème suivant d'Abel (1824) : Les équations de degré 5 ne sont pas résolubles au moyen des seules opérations d'addition, de soustraction, de multiplication, de division et d'extraction des racines n-ièmes effectuées sur leurs coefficients.

Puis, en étudiant les groupes linéaires sur un corps fini K, on découvrit d'autres groupes simples. Pour chaque entier $n \geqslant 1$, les $n \times n$ matrices (a_{ij}) à coefficients a_{ij} dans K, et de déterminant det (a_{ij}) non nul, forment un groupe $GL(n,K)$ pour la multiplication $(a_{ij})(b_{ij}) = (c_{ij})$ où :

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

avec $i = j = 1, ..., n$.

Le groupe $GL(n,K)$ s'appelle le *groupe linéaire général* de degré n sur K. L'application $(a_{ij}) \mapsto \det(a_{ij})$ est un morphisme surjectif du groupe $GL(n,K)$ sur le groupe multiplicatif du corps K. Le noyau $SL(n,$

K) de ce morphisme surjectif est donc un sous-groupe distingué de $GL(n, K)$. Le *centre* Z de $GL(n, K)$, formé des matrices de la forme xI, où I est la matrice identité et x un élément non nul de K, est aussi un sous-groupe distingué. Le groupe quotient $GL(n, K)/Z$ se note par $PGL(n, K)$, et l'image $SL(n, K)Z/Z$ de $SL(n, K)$ dans ce groupe par $PSL(n, K)$. Jordan et Dickson ont montré que, si $n \geqslant 3$, ou si $n = 2$ et si K a plus de trois éléments, le groupe $PSL(n, K)$ est simple ; c'est le seul facteur de Jordan-Hölder de $GL(n, K)$ qui soit simple et non cyclique.

La théorie des groupes linéaires généraux a été étendue aux autres groupes classiques par Jordan, Dickson et d'autres, vers la fin du XIXe siècle. Pour chaque groupe linéaire général, orthogonal, symplectique ou unitaire, ils trouvérent un ou plusieurs groupes correspondants (cela pour tout corps fini K). À quelques exceptions près (elles sont en nombre fini), tous ces groupes ont la propriété du groupe $GL(n, K)$; ils ont un et un seul facteur de Jordan-Hölder qui soit simple et non cyclique.

Les groupes simples définis par les groupes classiques sont eux-mêmes des cas spéciaux des groupes simples de Lie. En plus de ces groupes, il y a cinq groupes de Lie simples exceptionnels. Dickson découvrit des groupes finis correspondant à certains de ces groupes exceptionnels, mais c'est Chevalley. en 1955, qui donna une méthode générale de construction des groupes finis simples, correspondant à n'importe quel groupe simple de Lie (cf. la partie E ci-après Groupes de Lie). Il découvrit ainsi de nouveaux groupes finis simples, correspondant aux autres groupes de Lie exceptionnels. Steinberg (1959) compléta le travail de Chevalley en construisant des variantes de ces groupes. Pour cela, il s'est inspiré de la construction des groupes unitaires comme des variantes des groupes linéaires généraux.

En 1959, la liste des groupes finis simples qui étaient connus contenait les groupes alternés A_n, les groupes de Chevalley, leurs variantes de Steinberg et cinq groupes isolés : les quatre groupes M_{11}, M_{12}, M_{23}, M_{24} de Mathieu et le stabilisateur d'un élément de M_{23}, qui s'appelle le *groupe* M_{22} *de Mathieu.* Les groupes de Mathieu mis à part, tous ces groupes appartenaient à de bonnes familles infinies.

Cet ordre relatif fut ébranlé par Suzuki, qui découvrit, en étudiant les groupes de Zassenhaus, une nouvelle famille infinie de groupes simples de permutations. Mais Rhee s'aperçut bientôt, en 1961, que l'on pouvait construire ces groupes de Suzuki avec une nouvelle variante de la construction de Chevalley, variante possible seulement pour certains corps finis. En appliquant cette méthode à d'autres groupes simples de Lie, Rhee trouva deux nouvelles familles infinies de groupes finis simples. Les groupes de Suzuki et de Rhee n'étant que des variantes des groupes de Chevalley, tout rentra dans l'ordre.

L'ordre ne dura pas longtemps. En classant les groupes finis simples possédant des 2-sous-groupes de Sylow abéliens, Janko (1966) découvrit un nouveau groupe simple isolé d'ordre 175 560. Depuis lors, de nombreux autres groupes finis simples, apparaissant comme isolés et exceptionnels (d'où leur nom de « groupes sporadiques ») ont été découverts : le nombre de ces groupes sporadiques, incluant les groupes de Mathieu, était de vingt-six en 1980.

Il est nécessaire ici de préciser ce que l'on entend par « découverte » d'un groupe fini simple sporadique. En effet, l'existence de nombre de ces groupes a

souvent été *prédite* avant d'être effectivement prouvée, la prédiction se fondant sur des coïncidences numériques ou structurelles que le hasard seul ne pouvait raisonnablement expliquer. On peut noter l'analogie de cette démarche scientifique avec certains aspects de l'histoire de la théorie des particules élémentaires. C'est ainsi que Janko a prédit en 1974 l'existence du quatrième des groupes sporadiques qui porte son nom, mais qu'il a fallu attendre 1980 pour qu'une construction effective de ce groupe soit menée à bien à l'université de Cambridge, à l'aide de nombreuses heures de calcul sur ordinateur.

Le plus « gros » des groupes sporadiques est celui que l'on a dénommé le « monstre », dont l'existence a été prédite en 1973 par B. Fischer et R. Griess ; son nombre d'éléments est de l'ordre de 10^{54}. Malgré l'énormité de ce nombre, il semble pourtant que l'existence des groupes sporadiques traduise le phénomène souvent rencontré que certaines propriétés des nombres entiers ne sont vraies qu'à partir d'une certaine valeur : les petits nombres sont souvent exceptionnels. De fait, le plus grand nombre premier divisant le nombre d'éléments d'un groupe sporadique est 71 (en l'occurrence, 71 divise l'ordre du monstre). L'existence des groupes sporadiques semble également liée aux coïncidences ou isomorphismes exceptionnels entre « petits » groupes classiques.

Les spécialistes ont annoncé en 1980 que la classification des groupes finis simples était achevée, et par conséquent que les seuls groupes finis simples sont les groupes cycliques d'ordre premier, les groupes alternés de degré supérieur à 5, les groupes classiques et leurs variantes construits par Chevalley et ses successeurs (groupes appelés désormais « groupes de type de Lie »), et enfin les vingt-six groupes sporadiques connus. Le nombre de pages imprimées consacrées au travail de classification des groupes finis simples est de l'ordre de 10 000 ; les démonstrations s'appuient très souvent les unes sur les autres, et sont en général très complexes. Aucun mathématicien n'est en mesure, aujourd'hui, de vérifier tout seul cette gigantesque démonstration ou d'en comprendre les mécanismes profonds, et certains spécialistes pensent donc que, dans son état actuel, elle n'est pas tout à fait convaincante. Il est à noter que cette démonstration ne fournit aucune preuve a priori de l'existence d'un nombre *fini* de groupes sporadiques. Ce n'est qu'a posteriori, une fois la classification établie, que l'on constate que tel est bien le cas.

Le principe de la démonstration est celui de la récurrence : considérant un groupe G dont tous les facteurs de Jordan-Hölder de tous les sous-groupes propres sont isomorphes à l'un des groupes « connus », on essaye de démontrer que G est lui-même isomorphe à un groupe connu. L'outil essentiel de cette démarche est l'étude des centralisateurs d'involutions, qui repose sur deux résultats fondamentaux. Le premier est dû à R. Brauer, et affirme qu'il n'y a qu'un nombre *fini* de groupes finis possédant une involution (c'est-à-dire un élément de carré l'identité) dont le centralisateur (c'est-à-dire l'ensemble des éléments du groupe qui commutent avec cette involution) soit isomorphe à un groupe fixé.

Le deuxième, dont la démonstration est l'une des plus difficiles de l'histoire des mathématiques, est dû à W. Feit et J. G. Thompson (1963) ; il affirme que *tous les groupes finis et simples sont d'ordre pair, à l'exception des groupes cycliques* C_p *d'ordre premier p*. Ce remarquable théorème a été une conjecture « impossible »

pendant plus de cinquante ans. Il est à la base de plusieurs travaux qui cherchent à classer certaines sous-familles de groupes simples distingués par les propriétés de leurs sous-groupes. Par exemple, les groupes finis simples dont les 2-sous-groupes de Sylow sont dièdres furent classés par Gorenstein et Walter (1965). De tous les théorèmes de classification, le plus important et le plus difficile est le théorème de Thompson (1966) pour les groupes simples *minimaux*, c'est-à-dire pour les groupes finis simples G tels que chaque sous-groupe $H \neq G$ soit résoluble. Cette classification a déjà permis de démontrer plusieurs conjectures anciennes, ainsi que de nouveaux théorèmes.

L'un des aspects les plus mystérieux de la théorie des groupes finis est le phénomène baptisé *moonshine* par J. H. Conway. Il s'agit d'une correspondance extraordinaire entre le monstre et la théorie des fonctions automorphes, correspondance *constatée* mais à ce jour totalement inexpliquée. Il existe une série formelle :

$$J(X) = \frac{1}{X} + \sum_{n \geqslant 1} C_n \cdot X^n,$$

à coefficients dans l'anneau des caractères du monstre, possédant la propriété suivante : pour tout élément g du monstre, la fonction $j(g)$ définie sur le demi-plan de Poincaré $\mathcal{H} = \{z \in \mathbf{C} / \mathrm{Im}\, z > 0\}$ par :

$$j(g)(z) = \frac{1}{q} + \sum_{n \geqslant 1} C_n(g) q^n, \quad \text{où } q = e^{2\pi i z}$$

est une fonction remarquable ; plus précisément, il existe un sous-groupe discret $\Gamma(g)$ de $\mathbf{GL}_2(\mathbf{Q})$ tel que la surface $\mathcal{H}/\Gamma(g)$ se compactifie en une surface de genre 0, et toutes les fonctions modulaires pour $\Gamma(g)$ s'expriment rationnellement en $j(g)$. La fonction $j(1)$ est, à une constante près, l'invariant modulaire j.

3. p-groupes

Si H est un sous-groupe d'un groupe fini G, son ordre $|H|$, son indice $[G : H]$ (c'est-à-dire le nombre de classes à gauche de H dans G) et l'ordre $|G|$ de G sont liés par le *théorème de Lagrange* (1770) :

$$|G| = [G : H] \, |H| . \tag{2}$$

En particulier, $|H|$ divise $|G|$. Soit p un nombre premier et p^n la plus grande puissance de p qui divise $|G|$. Tout *p-sous-groupe* H de G (c'est-à-dire tout sous-groupe dont l'ordre $|H|$ est une puissance p^k de p) a un ordre $|H| \leqslant p^n$. Si $|H| = p^n$, on dit que H est un *p-sous-groupe de Sylow* de G. Il y a plusieurs théorèmes de Sylow (1872) pour ces sous-groupes :

1. Tout groupe fini G a au moins un p-sous-groupe de Sylow P.

2. Tout autre p-sous-groupe de Sylow Q de G est un conjugué de P, c'est-à-dire que $Q = \sigma P \sigma^{-1}$, pour un élément σ de G.

3. Tout p-sous-groupe H de G est un conjugué d'un sous-groupe de P, c'est-à-dire que $\tau H \tau^{-1} \subset P$, pour un élément τ de G.

4. Le nombre l des p-sous-groupes de Sylow de P divise $|G|$ et est de la forme $l = 1 + pm$, pour un certain entier m.

Le deuxième de ces théorèmes implique que le p-sous-groupe de Sylow P est déterminé, à un isomorphisme près, par le groupe G. On peut donc classer les groupes finis suivant leurs p-sous-groupes de Sylow ; d'où l'importance de la théorie des p-groupes (groupes P dont l'ordre est une puissance $p^k > 1$ d'un nombre premier p). L'une des propriétés de ces groupes est que *leurs centres* Z(P) *sont toujours non*

triviaux, soit Z(P) $\neq \{1\}$. *Chaque p-groupe* P *est donc nilpotent* (cf. la partie A ci-dessus - Généralités, fin du chap. 3).

Si G est un groupe fini et si E est un sous-ensemble de G, le *normalisateur* No(E) *de* E *dans* G est le sous-groupe formé des éléments σ de G, tels que $\sigma E \sigma^{-1} = E$. On montre alors qu'un groupe fini G est nilpotent si et seulement si tout sous-groupe H de G, différent de G, est strictement contenu dans son normalisateur No(H). On peut aussi montrer qu'un groupe fini G est nilpotent si et seulement si G n'a qu'un p-sous-groupe de Sylow G_p pour chaque nombre premier p. Dans ce cas, G est le produit direct de ses uniques sous-groupes de Sylow G_p. Les groupes finis nilpotents sont donc « presque » des p-groupes.

Plusieurs théorèmes relient la structure des normalisateurs No(H) des *p*-sous-groupes H $\neq \{1\}$ d'un groupe fini G avec celle de G. Frobenius a, par exemple, donné un critère pour la *p-nilpotence de G*, c'est-à-dire pour l'existence d'un sous-groupe distingué K dans G tel que P $\cap$ K = $\{1\}$ et PK = G pour tout psous-groupe de Sylow P. Voici le critère : *un groupe fini* G *est p-nilpotent si le normalisateur* No(H) *est p-nilpotentpour toutp-sous-groupe* H $\neq \{1\}$ *de* G. Notons qu'en général ces normalisateurs No(H) sont plus petits que G (par exemple, si G est simple et non cyclique).

Certains travaux récents de Thompson ont montré qu'il n'est pas nécessaire de vérifier lap-nilpotence de N,(H) pour *tout* p-sous-groupe H de G. Il suffit d'en choisir quelques-uns qui soient significatifs. Le plus important de ses résultats est le suivant : Soit P $\neq \{1\}$ un p-sous-groupe de Sylow de G, et soit *s* le maximum des ordres $|S|$ des sous-groupes *commutatifs* S de P. L'intersection A des sous-groupes commutatifs S de P, ayant pour ordre $|S| = s$, est un sous-groupe non trivial et distingué de P. Thompson a montré que *le groupe* G *est p-nilpotent si* No(A) *l'est et si* $p \geqslant 3$. Il suffit donc de regarder le normalisateur du seul p-sous-groupe A de G.

Il y a d'autres généralisations du théorème de Frobenius à des théorèmes sur l'existence de p-groupes quotients du groupe G sous certaines conditions sur les normalisateurs des p-sous-groupes de G. Les résultats de ce type jouent un rôle important dans l'étude des groupes simples et, en particulier, dans le théorème de Feit et Thompson.

EVERETT DADE

Bibliographie

N. BLANCKBURN & B. HUPPERT, *Finite Groups*, Springer, New York-Berlin, 1981 / J. CALAIS, *Éléments de théorie des* groupes, P.U.F.. Paris, 1984 / W. FEIT, *Representation Theory of Finite Groups*, Elsevier Science, New York, 1982 / D. GORENSTEIN, *The Classification of Finite Simple Groups*, Plenum, 1983 / M.-P. MALLIAVIN, J.-P. BÉZIVIN & A. LÉVY-BRUHL, *Les Groupes finis et leurs représentations complexes*, Masson, 1981-1982.

D. Représentation linéaire des groupes

Développée d'abord comme moyen de classification des différentes apparences du même groupe G comme groupe de transformations linéaires, la théorie des représentations linéaires est devenue un des outils les plus puissants pour l'étude de la structure de G. En particulier, les caractères irréductibles d'un groupe fini G, introduits pour mieux classer les représentations linéaires, sont vitaux pour la théorie moderne des groupes simples.

1. Représentation des groupes

À chaque système mathématique S est associé son groupe de symétries (ou d'automorphismes) X(S). On considère ces groupes $\Sigma(S)$ comme étant concrets. Une *représentation* R d'un groupe quelconque G *comme groupe de symétries de S* est un homomorphisme $\sigma \mapsto R_\sigma$ de G dans le groupe concret C(S). Elle donne une réalisation de la loi de composition abstraite de G comme loi de composition concrète dans C(S).

La théorie des représentations cherche les conséquences, pour les deux structures S et G, de l'existence d'une représentation R, et les utilise pour démontrer des théorèmes qui n'ont quelquefois rien à voir avec les représentations ; par exemple, le théorème de Feit et Thompson : tout groupe d'ordre impair est résoluble. Seules les relations entre S et G étant intéressantes, la tendance moderne est de les définir directement et de supprimer le groupe C(S) et l'homomorphisme R. Voici, sur un exemple, comment on procède.

Une *opération* d'un groupe G sur un ensemble E est une loi de composition externe, envoyant tout élément σ de G et tout élément x de E sur un élément σx de E, et suppose que cette loi satisfait aux conditions :

(1 *a*) $$1x = x$$

pour tout x dans E,

(1 *b*) $$(\sigma\tau)x = \sigma(\tau x),$$

pour tout σ, τ dans G et tout x dans E, où $\sigma\tau$ est le produit dans G, et 1 l'élément neutre de G. Ces conditions impliquent que, pour tout σ de G, l'application R_σ : $x \mapsto \sigma x$ est bijective de E sur E, c'est-a-dire que R_σ est une permutation sur l'ensemble E. Et l'application $\sigma \mapsto R_\sigma$ est un homomorphisme de G dans le groupe C(E) des permutations sur E (que l'on peut considérer comme les symétries de E). L'opération de G sur E détermine donc une représentation R de G comme groupe de symétries de E. En fait, la représentation et l'opération ne sont que deux façons de voir la même chose, car la première détermine la seconde par la relation $\sigma x = R_\sigma(x)$, pour tout σ de G et tout x de E.

Une *représentation linéaire* d'un groupe G est une représentation de G comme groupe de symétries d'un espace vectoriel. Rappelons qu'un *espace vectoriel* V sur un corps K est un groupe additif muni d'une loi de composition externe, qui envoie tout élément λ de K et tout élément v de V sur un élément Av de V, et qui est telle que les combinaisons linéaires $\lambda_1 v_1 + + \lambda_n v_n$ d'éléments $v_1, \dots, v_n$ de V à coefficients $\lambda_1, \dots$ A, dans K obéissent aux règles ordinaires de calcul. Une *opération linéaire* de G sur V est une opération de G sur l'ensemble V, satisfaisant à la condition de linéarité :

$$\sigma(\lambda_1 v_1 + \quad + \lambda_n v_n) = \lambda_1 \sigma(v_1) + \quad + \lambda_n \sigma(v_n)$$

pour tout σ de G, tout A,, ... λ_n de K, et tout $v_1, \dots, v_n$ de V. Chaque application R_σ : v - σ est alors une *transformation Maire bijective* de V sur lui-même, et l'homomorphisme $\sigma \mapsto$ R', est une *représentation linéaire* de G sur V. On dit alors que V est un *G-espace.*

On dit que deux représentations $\sigma \mapsto R_\sigma$ et $\sigma \mapsto R_{\sigma'}$ sur des espaces vectoriels V et V′ sont *équivalentes* (ou isomorphes) s'il existe un isomorphisme linéaire de V sur V' tel que $R_\sigma = \varphi^{-1} \circ R'_\sigma \circ \varphi$ pour tout $\sigma \in G$, ce qui équivaut à $\varphi \circ R_\sigma = R'_\sigma \circ \varphi$.

Il y a une autre façon, souvent utile, de considérer les représentations linéaires. Supposons que l'espace vectoriel V ait une base finie $v_1, \dots, v_d$, c'est-à-dire que tout élément v de V ait une expression unique : $v = \lambda_1 v_1 + \dots + \lambda_d v_d$, comme combinaison linéaire des éléments $v_1, \dots, v_d$ de la base, à coefficients $\lambda_1, \dots$ A, dans K. Pour tout élément σ de G, il existe alors des éléments uniques $\alpha_{ij}(\sigma)$ dans K tels que :

$$\sigma v_j = \sum_{i=1}^{d} a_{ij}(\sigma) v_i, \quad \text{pour } j = 1, \dots, d.$$

La d X d matrice, A(o) = (a,,(o)), détermine la transformation linéaire R_σ par les équations ci-dessus. Les conditions (1) équivalent à dire que l'application *A :* $\sigma \mapsto$ *A(o)* est un *homomorphisme du groupe G dans le groupe* GL(d, K) de toutes les matrices d X d dont le déterminant n'est pas nul.

On peut donc regarder les représentations linéaires ou les opérations linéaires comme des homomorphismes dans GL(d, K).

Cette façon de voir facilite la définition du *caractère* χ_V du G-espace V. C'est la fonction de G dans K dont la valeur $\chi_V(\sigma)$ est la *trace,* Tr(A(o)), de la matrice *A(a) :*

$$\chi_V(\sigma) = \mathrm{Tr}(A(\sigma)) = a_{11}(\sigma) + a_{22}(\sigma) + \dots + a_{dd}(\sigma),$$

pour tout σ de G.

La trace de *A(o)* est indépendante du choix de la base $v_1, \dots, v_d$ et ne dépend que de la transformation linéaire Ro. Le caractère χ_V est donc déterminé par l'opération. Dans certains cas importants, la réciproque est vraie, c'est-à-dire que l'opération de G sur V est déterminée par le caractère χ_V (cf. *infra*).

11 faut noter la propriété suivante du caractère χ_V, qui découle de l'équation $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ pour des matrices A et B :

(2) $$\chi_V(\tau^{-1}\sigma\tau) = \chi_V(\sigma),$$

pour tout σ, τ dans G.

On dit que deux éléments, σ et ρ de G sont conjugués s'il existe un élément τ de G tel que $\rho = \tau^{-1}\sigma\tau$. Les classes d'équivalence pour cette relation de conjugaison s'appellent les classes de conjugaison de G. La condition (2) signifie donc que chaque caractère χ_V est constant sur chaque classe de conjugaison de G.

2. Théorie des représentations linéaires d'un groupe fini

La théorie classique trouvée par G. Frobenius, W. Burnside, et I. Schur dans la période 1890-l 9 10 est la base de **toutes** les généralisations modernes. Cette théorie s'applique **aux** représentations linéaires d'un groupe fini G sur des espaces vectoriels de dimensions finies (c'est-à-dire ayant une base finie) sur le corps C des nombres complexes.

On cherche, d'abord, à classer les G-espaces, à des isomorphismes près. Un G-espace V est *G-isomorphe à* un G-espace U s'il existe une application linéaire bijectivefde V sur U qui conserve les opérations de G : $f(\sigma v) = \sigma f(v)$, pour tout σ dans G, et tout v dans V, c'est-à-dire si les représentations linéaires sont équivalentes.

L'outil principal de la classification des opérations linéaires de G sur des espaces vectoriels V est la décomposition des espaces V en somme directe de sous-espaces stables. Un *sous-espace* de V est un sous-ensemble U, qui est fermé pour la

formation de combinaisons linéaires d'éléments ; U est donc lui-même un espace vectoriel avec, pour lois de composition, les restrictions des lois de composition de V. Le sous-espace U est *stable* par G s'il est fermé pour l'opération de G sur U, c'est-à-dire si ou appartient à U pour tout σ dans G et tout u dans U. Dans ce cas, la restriction à U de l'opération de G sur V est une opération linéaire de G sur U. Soit U,, ... U_k des sous-espaces stables de V. On dit que V est la *somme directe* $U_1 \oplus ... \oplus U_k$ des U_i, si tout élément v de V a une expression unique de la forme :

$$v = u_1 + u_2 + ... + u_k,$$

où u_i appartient à U_i pour $i = 1, ... k$.

Les éléments $u_1, ... u_k$ sont les *composantes* de v pour la décomposition V = $U_1 \oplus ... \oplus U_k$. La correspondance :

$$v \leftrightarrow (u_1, ..., u_k)$$

est une bijection de V sur le produit cartésien U_1 X . . X U_k ; les lois de composition de V et l'opération de G sur V se calculent à partir des structures des U_i par les relations :

$$v + Y' = (u_1 + u'_1) + ... + (u_k + u'_k),$$
$$\lambda = (\lambda u_1) + + (\lambda u_k),$$
$$\sigma v = (\sigma u_1) + + (\sigma u_k),$$

où λ appartient au corps, σ au groupe, et où v' est un élément de V ayant comme composante les éléments $u'_1, ... u'_k$. La décomposition V = U, $\oplus \oplus U_k$ donne donc une analyse de la structure de V au moyen de celles des U_i.

Si G opère sur un espace vectoriel U $\neq \{0\}$, et s'il n'y a aucun sous-espace stable par G, sauf U et $\{0\}$, on dit que le G-espace U est *irréductible.* La classification des opérations linéaires d'un groupe fini G sur des espaces vectoriels V de dimension finie sur le corps C des nombres complexes est contenue dans les deux énoncés suivants :

(3*a*) l'espace V a au moins une décomposition : V = $U_1 \oplus \oplus U_k$, en somme directe de sous-espaces stables et irréductibles $U_1, ... U_k$;

(3b) si V = $U'_1 \oplus \oplus$ U', est une autre telle décomposition, alors $l = k$ et, après une permutation convenable des indices, U_i est G-isomorphe à U'_i, pour $i = 1, ... k$.

Pour tout G-espace irréductible W, on définit la *multiplicité,* m(W dans V), de W dans V. C'est le nombre des indices $i = 1, ... l$ pour lesquels W est G-isomorphe à U,. À cause de (3 *b),* cette multiplicité est indépendante de la décomposition V = $U_1 \oplus \oplus U_k$. Donc deux G-espaces V et V' sont isomorphes si, et seulement si :

$$m\,(\text{W dans V}) = m\,(\text{W dans V}'),$$

pour tout G-espace irréductible W.

Frobenius découvrit une méthode très simple de calcul des multiplicités m(W dans V) en utilisant les caractères. On définit un produit hermitien $(f|g)_G$ sur l'espace vectoriel Fct(G, C) de toutes les fonctions de G dans C par :

$$(f|g)_G = \frac{1}{|G|} \sum_{\sigma \in G} f(\sigma)\overline{g(\sigma)},$$

pour tout f et g dans Fct(G, C), ou $\overline{g(o)}$ désigne le complexe conjugué de g(o) et $|G|$ est *l'ordre* de G, c'est-à-dire le nombre d'éléments dans le groupe fini G. Si W et U sont deux G-espaces irréductibles de dimension finie sur G, Frobenius a démontré *les relations d'orthogonalité* suivantes, pour leurs caractères χ_W et χ_U :

(4 *a*) $$(\chi_W | \chi_U)_G = 1,$$

si W est G-isomorphe à U ;

(4 *b*) $(\chi_W \quad \chi_U)_G = 0,$

si W n'est pas G-isomorphe à U.

La décomposition, $V = U_1 \oplus \quad \oplus U_k$, entraîne pour les caractères la relation :

$$\chi_V = \chi_{U_1} + ... + \chi_{U_k}.$$

On a donc, d'après les relations d'orthogonalité (4), la formule suivante pour la multiplicité de W dans V :

$$m\,(W \text{ dans } V) = (\chi_W \quad \chi_V)_G$$

On voit que le caractère χ_V détermine les multiplicités m(W dans V). Le G-espace V est donc déterminé à un isomorphisme près par son caractère χ_V.

Les relations d'orthogonalité (4) montrent que les *caractères irréductibles* (les caractères des G-espaces irréductibles) distincts sont des fonctions linéairement indépendantes sur le groupe fini G. Il n'y a donc qu'un nombre fini $\chi_1, \dots, \chi_c$ de tels caractères. On peut alors montrer que le nombre c des caractères irréductibles de G est égal au nombre des classes de conjugaison de G.

Soit H un sous-groupe de G. Tout G-espace V est, par restriction, un H-espace V_H. Le caractère de cet H-espace est la restriction χ_H du caractère χ du G-espace V au sous-groupe H. Donc la restriction $\chi \mapsto \chi_H$ est une application des caractères de G dans ceux de H. Frobenius découvrit une application, allant en sens inverse, envoyant tout caractère φ de H sur le *caractère induit* φ^G de G défini par :

(5) $$\varphi^G(\sigma) = \frac{1}{|H|} \sum_{\substack{\tau \in G \\ \tau^{-1}\sigma\tau \in G}} \varphi(\tau^{-1}\sigma\tau),$$

pour tout σ de G, où la sommation sur l'ensemble vide est nulle par définition. Frobenius a démontré la *lui de réciprocité :*

(6) $$(\chi_H | \varphi)_H = (\chi | \varphi^G)_G,$$

pour tout caractère χ de G et tout caractère φ de H.

On peut calculer les caractères irréductibles de G en utilisant les caractères induits : on part de certains sous-groupes H de G, et de certains caractères connus φ de H. À l'aide de (5), on calcule les caractères induits φ^G, qui sont, comme tous les caractères, des combinaisons linéaires $a_1\chi_1 + \quad + a_c\chi_c$ à coefficients entiers $a_1, \dots, a_c$ des caractères irréductibles $\chi_1, \dots, \chi_c$ de G. On cherche alors un nombre suffisant de caractères φ^G, de telle manière qu'ils engendrent le groupe additif X(G) de toutes ces combinaisons linéaires de $\chi_1, \dots, \chi_c$. Il faut donc trouver les éléments :

$$\chi = a_1\chi_1 + ... + a_c\chi_c$$

de X(G) tels que :

$$(\chi \quad \chi)_G = a_1^2 + \quad + a_c^2 = 1.$$

Un tel élément est forcément de la forme $\pm\chi_i$, où $i = 1, \dots, c$. Comme $\chi_i(1) > 0$, on peut déterminer le caractère χ_i.

Cette méthode est justifiée, car R. Brauer a montré que X(G) est en fait engendré par les caractères induits φ^G, où φ parcourt la famille de tous les caractères *linéaires* (c'est-a-dire irréductibles de degré 1) des sous-groupes H de G. On peut méme se restreindre aux sous-groupes H qui sont nilpotents. Un caractère linéaire φ d'un sous-groupe H est simplement un homomorphisme $\sigma \mapsto \varphi(\sigma)$ de H dans le groupe multiplicatif du corps C. Ces caractères sont donc faciles à calculer.

3. Les généralisations

La théorie classique, exposée ci-dessus, a été au fil des années généralisée de plusieurs façons. L'une d'elles consiste à remplacer le corps C des nombres complexes par un autre corps K. Si le corps K est de caractéristique zéro, ou p (l'entier p étant un nombre premier qui ne divise pas l'ordre fini $|G|$ de G), la théorie des représentations linéaires de G sur les espaces vectoriels de dimension finie sur K se réduit facilement à la théorie classique des caractères complexes de G, et l'on n'obtient aucune notion nouvelle. Par contre, si la caractéristique p de K est un nombre premier qui divise l'ordre $|G|$, on trouve une nouvelle famille de représentations irréductibles et de caractères de G, les *caractères modulaires.* L'étude de ces caractères modulaires et de leurs relations avec les caractères complexes, due surtout à R. Brauer, a permis de trouver, pour ces derniers, des lois et identités nouvelles. Plusieurs théorèmes importants sur les groupes simples n'ont pu être démontrés que grâce à la théorie de ces caractères.

Une autre famille de généralisations de la théorie classique concerne les représentations unitaires continues d'un groupe topologique sur un espace de Hilbert. Un *groupe topologique* G est un groupe muni d'une topologie par rapport à laquelle la multiplication et l'inversion sont des applications continues. Un *espace hilbertien* V est un espace vectoriel sur les nombres complexes C muni d'un produit hermitien $(u|v)$ (c'est-à-dire une application de V X V dans C telle que l'application $u \mapsto (u|v)$ est linéaire pour tout v dans V, $(u|v) = \overline{(v|u)}$ pour tout u et v dans V, et $(u|u)$ est un nombre réel strictement positif pour tout $u \neq 0$ dans V) et complet pour la norme $\|v\| = (v|v)^{1/2}$ définie par ce produit hermitien. Une opération linéaire de G sur V est *continue* si l'application $(\sigma, u, v) \rightarrow (\sigma v | u)$ est continue en tant qu'application de G X V × V dans *C.* Elle est *unitaire* si elle conserve le produit hermitien $(\sigma v | \sigma v) = (u|v)$ pour tout σ dans G et tout u, v dans V. On dit, dans ce cas, que V est un *G-espace de Hilbert.*

Lorsque le groupe topologique G est *compact,* la théorie est très semblable à la théorie classique. L'espace V admet alors une décomposition en somme orthogonale :

$$V = \bigoplus_{U \in F} U$$

d'une famille F de sous-espaces stables et irréductibles U. C'est-à-dire que les U sont des sous-espaces fermés de V, deux à deux orthogonaux pour le produit hermitien $(u|v)$, et tout élément v de V admet une décomposition unique :

$$v = \sum_{U \in F} v_U,$$

en une somme convergente en norme de ses composantes v_U appartenant à U. Toute autre telle décomposition :

$$V = \bigoplus_{W \in E} W$$

est équivalente à la première, en ce sens qu'il existe une application bijective f de E sur F et, pour tout W de E, une application linéaire bijective g de W sur $U = f(W)$ qui conserve les opérations de G et les produits hermitiens sur les deux sous-espaces W et U. Donc V est déterminé à un isomorphisme près par les *multiplicités* des G-espaces de Hilbert irréductibles W dans V, c'est-à-dire le nombre de sous-espaces U

appartenant à F, tels que U soit G-isomorphe à W.

On peut montrer que toute représentation unitaire continue irréductible d'un groupe compact est de dimension finie. Un groupe compact G possède donc des *caractères irréductibles*. Ces caractères satisfont aussi aux relations d'orthogonalité (4), où le produit hermitien $(\chi \quad \varphi)_G$ est défini par :

$$(\chi \mid \phi)_G = \int_G \chi(\sigma)\,\overline{\phi(\sigma)}\,d\sigma,$$

l'intégrale étant prise par rapport à la mesure de Haar normalisée (cf. **INTÉGRATION ET MESURE,** chap. 3) :

$$\int_G 1\,d\sigma = 1.$$

Lorsque le groupe G n'est pas compact, la théorie est beaucoup moins nette. Au lieu de décomposer V en sommes orthogonales, il faut le décomposer en *intégrales orthogonales* de G-espaces irréductibles. De telles décompositions, quand elles existent, ne sont pas nécessairement uniques. Les représentations irréductibles peuvent être de dimension infinie, et donc ne pas avoir de caractères. En général, tout devient très compliqué. Il y a néanmoins une théorie assez bonne pour les représentations des groupes classiques qui sont importants en mécanique quantique.

4. Applications aux groupes finis

Les caractères irréductibles $\chi_1, ..., \chi_c$ d'un groupe fini G forment un outil très puissant dans l'étude de G. On considère leurs valeurs comme des invariants numériques de G, invariants qui doivent satisfaire à plusieurs conditions fortes, comme les relations d'orthogonalité, et qui sont liés à la structure algébrique de G. On combine ces conditions et ces relations pour montrer des théorèmes parfois surprenants sur G.

On utilise d'abord les caractères pour trouver des sous-groupes distingués de G. Pour $i = 1, ..., c$, soit W_i un G-espace ayant χ_i pour caractère ; on appelle *noyau* de χ_i, le sous-groupe distingué $\mathrm{Ker}(\chi_i)$ formé de tous les σ de G opérant *trivialement* sur W, par $\sigma w = w$, pour tout w dans W_i. Il est important de noter que ce sous-groupe distingué est caractérisé par les valeurs de χ_i : c'est l'ensemble de tous les éléments σ de G tels que $\chi_i(\sigma) = \chi_i(1)$.

Il n'y a qu'un seul caractère χ_i tel que $\mathrm{Ker}(\chi_i) = G$, le *caractère trivial* χ_1 dont les valeurs sont $\chi_1(\sigma) = 1$ pour tout σ de G. Pour **tout** caractère non trivial χ_i, $i \geqslant 2$, il existe un élément σ de G tel que $\chi_i(\sigma) \neq \chi_i(1)$.

On peut aussi montrer que, pour tout élément $\sigma \neq 1$ de G, il existe au moins un caractère non trivial χ_i tel que $\chi_i(\sigma) \neq \chi_i(1)$.

Une connaissance très grossière des valeurs des caractères de G peut permettre de prouver l'existence d'un sous-groupe distingué K qui soit non trivial ($K \neq \{1\}$, $K \neq G$). Il suffit de trouver un seul élément $\sigma \neq 1$ et un seul caractère non trivial χ_i tel que $\chi_i(1) = \chi_i(\sigma)$. Le sous-groupe $K = \mathrm{Ker}(\chi_i)$ est alors un sous-groupe distingué non trivial.

On s'intéresse aux relations entre la structure des sous-groupes de G et les caractères de G. Une de ces relations a trait aux ensembles à intersections triviales. Un tel ensemble S est un sous-ensemble d'un sous-groupe H, appelé *normalisateur* de S, dont les conjugués $\sigma^{-1}S\sigma$ satisfont à :

$$(7a) \qquad S = \sigma^{-1}S\sigma,$$

si σ appartient à H ;

(7 b) $$S \cap (\sigma^{-1} S \sigma) = \varnothing$$

si σ est dans Ci mais non dans H.

Voici un exemple d'un tel ensemble S : pour tout σ de G, on désigne par Rac(σ) l'ensemble de tous les τ de G qui sont racines de CI (il existe n tel que $\tau^n = a$). Rac(σ) est un ensemble à intersections triviales, et le sous-groupe H est le normalisateur du sous-groupe cyclique $< \sigma >$ engendré par σ. C'est le groupe formé de tous les éléments ρ de G, tels que $\rho^{-1} < \sigma > \rho = < \sigma >$.

Soit $\varphi_1, \dots, \varphi_c$ les caractères irréductibles de H. On désigne par X(H S) le groupe additif formé des combinaisons linéaires :

$$\psi - a_1\varphi_1 + + a_c\varphi_c$$

(à coefficients entiers a_i) qui s'annulent hors de S. La fonction induite ψ^G, définie par (5), appartient alors au groupe X(G) formé des combinaisons linéaires à coefficients entiers des $\chi_1, \dots, \chi_c$. Les conditions (7) et la formule (5) entraînent :

(8) $$\psi^G(\sigma) = \psi(\sigma),$$

pour tout σ dans S et tout ψ dans X(H S).

En combinant cette équation avec la loi de réciprocité (6), on trouve que l'application $\psi \mapsto \psi^G$ est une *isométrie* X(H S) $\mapsto$ X(G), c'est-à-dire que l'on a :

(9) $$(\varphi^G | \psi^G)_G = (\varphi | \psi)_H,$$

pour tout φ et ψ dans X(H| S).

Frobenius fut le premier à utiliser cette isométrie. Il considéra le cas où S est égal au sous-groupe H moins l'élément neutre 1. On désigne par K le sous-ensemble de tous les éléments τ de G qui n'appartiennent à aucun conjugué σ 'Sa de S. Le théorème de Frobenius dit que l'ensemble K est un sous-groupe distingué de G. On appelle groupe de Frobenius tout groupe G possédant un sous-groupe H différent de { l} et de G, ayant la propriété énoncée dans le théorème de Frobenius. Le sous-groupe distingué K s'appelle le noyau *de Frobenius* (cf. la partie C ci-dessus Groupes finis).

La théorie des caractères exceptionnels est basée sur l'isométrie (9). Supposons que φ_i et φ_j soient deux caractères irréductibles distincts de H tels que φ_i φ_j appartienne à X(H S). On a alors :

$$(\varphi_i - \varphi_j)^G = a_1\chi_1 + \dots + a_c\chi_c$$

pour certains entiers a,, ... a_c. L'isométrie (9) et les relations d'orthogonalité (4) donnent :

$$\begin{aligned} 2 &= (\varphi_i - \varphi_j | \varphi_i - \varphi_j)_H \\ &= ((\varphi_i - \varphi_j)^G | (\varphi_i - \varphi_j)^G) = a_1^2 + \dots + a_c^2. \end{aligned}$$

Mais les a_i sont des entiers. Ils sont donc tous nuls, sauf deux d'entre eux, a, et a_j, qui valent ± 1. Si $a_i = a_j = \pm 1$, on a :

$$0 = (\varphi_i - \varphi_j)^G(1) = \pm(\chi_i(1) + \chi_j(1)),$$

ce qui est impossible, car les degrés $\chi_i(1)$ et $\chi_j(1)$ sont des entiers strictement positifs. On a donc :

$$(\varphi_i - \varphi_j)^G = \pm(\chi_i - \chi_j).$$

On dit que χ_i et χ_j sont les *caractères exceptionnels* de G correspondant à φ_i et φ_j. À cause de (5), ces deux caractères sont égaux en dehors des conjugués de S. Sur ces conjugués, leur différence est déterminée par (8). Ce type de résultat intervient, par exemple, dans la démonstration par Feit et Thompson de leur célèbre théorème : « Tout groupe d'ordre impair est résoluble. » Feit et Thompson utilisent la théorie des caractères excep-

tionnels pour des isométries qui généralisent (9).

EVERETI DADE

Bibliographie

R. BRAUER, *Collected Papers, 3 vol.*, W. J. Warren et P. Fong éd.. M.I.T. Press, Cambridge (Mass.), 1980 / C. W. CURTIS & I. REINER, *Methods of Representation Theory*, J. Wiley, New York, 1990 / J.-P. SERRE, *Représentations linéaires des groupes finis*, Hermann, Paris, 3' éd. 1978 / E. B. VINBERG, *Linear Representations of Groups*, Birkhauser Boston. Cambridge (Mass.). 1989 / H. WEYL, *The Theory of Groups and Quantum Mechanics*, Dover. New York. 1950.

E. Groupes de Lie

La théorie des groupes de Lie, fondée dans la période de 1870-1880 par le mathématicien norvégien Sophus Lie, a d'abord été considérée comme une partie assez marginale des mathématiques, liée à des problèmes touchant les équations différentielles, les équations aux dérivées partielles et la géométrie différentielle. Leur étude générale a mis plus tard en évidence un certain nombre d'objets mathématiques particuliers, explicitement définis, les groupes *semi-simples*, dont on a peu à peu découvert le rôle fondamental dans presque toutes les parties des mathématiques modernes, même les plus éloignées en apparence des vues initiales de Lie. En outre, ces groupes semblent intervenir de façon de plus en plus profonde dans les conceptions récentes de la physique théorique, surtout en théorie de la relativité et en mécanique quantique.

On suppose connues les notions fondamentales relatives aux variétés différentielles et analytiques. On utilisera systématiquement ici les notions introduites dans l'article sur les groupes classiques, qui constituent les premiers et les plus importants exemples de groupes de Lie.

1. La structure des groupes de Lie généraux

Un *groupe de Lie* (appelé aussi *groupe de Lie réel*) est, par définition, une variété analytique réelle G (dite *sous-jacente* au groupe), munie d'une loi de composition $(x, y) \mapsto xy$ pour laquelle G est un groupe, et qui est telle que l'application $(x, y) \mapsto xy^{-1}$ de G X G dans G soit analytique. Une variété analytique *complexe G* munie d'une loi de composition $(x, y) \mapsto xy$ pour laquelle G est un groupe, et qui est telle que $(x, y) \mapsto xy^{-1}$ soit une application *holomorphe* de G X G dans G, est appelée *groupe de Lie complexe* ; un tel groupe peut évidemment aussi être considéré comme groupe de Lie *réel* (dit *sous-jacent* au groupe de Lie complexe), en n'envisageant que sa structure de variété analytique réelle. Dans un groupe de Lie réel (resp. complexe) G, les translations $x \mapsto ax$, $x \mapsto xa$ et les automorphismes intérieurs :

$$\mathrm{Int}(a) : x \mapsto axa^{-1}$$

sont des applications analytiques (resp. holomorphes) ; il en résulte que la dimension (resp. la dimension complexe) de la variété sous-jacente à G est la même en tous les points de G ; on dit que c'est la *dimension* (resp. la *dimension complexe*) de G ; si G est un groupe de Lie complexe de dimension complexe n, le groupe de Lie réel sous-jacent est de dimension $2n$.

Un *sous-groupe de Lie* (resp. sous-groupe de Lie *complexe)* H d'un groupe de Lie (resp. groupe de Lie complexe) G est un sous-groupe de G dont l'ensemble sous-jacent est une *sous-variété fermée* de la

variété sous-jacente à G. On montre qu'un *sous-groupe fermé* d'un groupe de Lie G est nécessairement un sous-groupe de Lie de G (mais non nécessairement un sous-groupe de Lie *complexe* lorsque G est un groupe de Lie complexe).

Ainsi, le groupe linéaire GL (n, R) (resp. $\mathbf{GL}(n, \mathrm{C})$) est un groupe de Lie réel (resp. complexe) de dimension (resp. de dimension complexe) n^2 ; le groupe unimodulaire $\mathbf{SL}(n, \mathrm{R})$ (resp. $\mathbf{SL}(n, \mathrm{C})$) en est un sous-groupe de Lie (resp. un sous-groupe de Lie complexe) de dimension (resp. de dimension complexe) n^2 1. Le groupe orthogonal O(n, R) (resp. O(n, C)) est un sous-groupe de Lie (resp. un sous-groupe de Lie complexe) de $\mathbf{GL}(n, \mathrm{R})$ (resp. $\mathbf{GL}(n, \mathrm{C})$ de dimension (resp. de dimension complexe) $n(n - 1)/2$.

Un homomorphisme $f : G \longrightarrow G'$ de groupes de Lie (resp. de groupes de Lie complexes) est un homomorphisme de groupes qui est en même temps une application analytique (resp. holomorphe). Le noyau $f^{-1}(e')$ est un sous-groupe de Lie (resp. un sous-groupe de Lie complexe) de G. Par contre, l'image f(G) est un sous-groupe de G' qui n'est pas nécessairement fermé.

On se bornera, dans ce chapitre, aux propriétés des groupes de Lie réels ; lorsqu'on mentionnera un groupe de Lie complexe, par exemple $\mathbf{GL}(n, \mathrm{C})$, c'est en fait le groupe de Lie réel sous-jacent dont il sera question.

Si N est un sous-groupe de Lie *distingué* d'un groupe de Lie G (donc fermé dans G), on peut définir sur le groupe G/N une structure et une seule de variété analytique qui en fait un groupe de Lie et pour laquelle l'application canonique :

$$\pi : G \rightarrow G/N$$

est une submersion. Si G et G' sont deux groupes de Lie, la structure de variété produit sur G X G' définit sur ce groupe une structure de groupe de Lie. Plus généralement, soit L et N deux groupes de Lie, et $x \mapsto \tau_x$ un homomorphisme du groupe L dans le groupe Aut(N) des automorphismes de N, tel que l'application $(x, y) \mapsto \tau_x(y)$ de L X N dans N soit analytique. Alors le produit semi-direct $L \times_\tau N$ de L par N relatif à τ est un groupe de Lie pour la structure de variété produit; N est un sous-groupe distingué fermé de $L \times_\tau N$, et le groupe quotient :

$$(L \times_\tau N)/N$$

est isomorphe à L.

Les groupes de Lie de dimension 0 sont les groupes *discrets*. Dans un groupe de Lie G, la composante neutre (c'est-à-dire la composante connexe de l'élément neutre e de G) est un sous-groupe ouvert (donc à plus forte raison fermé) distingué G^0, et le quotient G/G^0 est discret ; on notera que G n'est pas nécessairement produit semi-direct de G^0 et d'un sous-groupe isomorphe à G/G^0. L'étude des groupes de Lie se concentre presque exclusivement sur les groupes de Lie *connexes*.

Pour un groupe de Lie connexe G, il existe un groupe de Lie *simplement* connexe $\tilde{G}$ appelé *revêtement universel* de G, déterminé à un isomorphisme près. tel que G soit isomorphe à $\tilde{G}/D$, où D est un sous-groupe discret du centre de $\tilde{G}$, isomorphe au groupe fondamental $\pi_1(G)$ de la variété sous-jacente à G (le groupe $\pi_1(G)$ est donc toujours *commutatif*). La structure des groupes de Lie connexes est donc ramenée à celle des groupes *simplement connexes*.

Ainsi le groupe $\mathbf{GL}(n, \mathrm{C})$ est connexe ; mais le groupe $\mathbf{GL}(n, \mathrm{R})$ a deux composantes connexes, la composante neutre étant l'ensemble des matrices de détermi-

nant strictement positif. Les groupes $\mathbf{SL}(n, \mathbf{C})$ et $\mathbf{SL}(n, \mathbf{R})$ sont connexes. Chacun des groupes O(n, C) et O(n, R) a deux composantes connexes ; les composantes neutres sont $\mathbf{SO}(n, \mathbf{C})$ et $\mathbf{SO}(n, \mathbf{R})$.

Un groupe de Lie commutatif connexe est nécessairement isomorphe à un groupe du type $\mathbf{R}^p \times \mathbf{T}^q$, où $\mathbf{T} = \mathbf{R}/\mathbf{Z}$ (« tore à une dimension ») ; son revêtement universel est $\mathbf{R}^{p+q}$. Le groupe $\mathbf{SL}(n, \mathbf{C})$ est simplement connexe, mais $\mathbf{GL}(n, \mathbf{C})$ ne l'est pas : son revêtement universel est isomorphe à :

$$\mathbf{SL}(n, \mathbf{C}) \times \mathbf{R}^2$$

et son groupe fondamental isomorphe à Z. Le groupe SO(2, R) est commutatif et isomorphe à T ; pour $n \geqslant 3$, le groupe $\mathbf{SO}(n, \mathbf{R})$ a pour revêtement universel le groupe $\mathrm{Spin}(n)$ noté encore $\mathrm{Spin}(n, \mathbf{R})$, et le groupe fondamental est d'ordre 2 ; on définit de même le groupe $\mathrm{Spin}(n, \mathbf{C})$, qui est revêtement universel de $\mathbf{SO}(n, \mathbf{C})$ pour $n \geqslant 3$, avec encore un groupe fondamental d'ordre 2.

Dans un groupe de Lie *simplement connexe* G, les groupes dérivés successifs D'(G) sont des sous-groupes distingués *fermés* connexes ; il en est de même des sous-groupes $C^r(G)$ de la série centrale descendante. Un groupe simplement connexe *résoluble* G a une variété sous-jacente isomorphe à un espace $\mathbf{R}^n$; son groupe dérivé D(G) est nilpotent, et il existe une suite croissante (H_j), $0 \leqslant j \leqslant n$, de sous-groupes fermés distingués de G telle que $H_0 = \{e\}$, $H_n = G$ et que H_{j+1}/H_j soit isomorphe à R. Par exemple, *le groupe trigonal large supérieur* $\mathrm{T}(n, \mathbf{R})$ formé des matrices réelles :

$$(x_{ij}) \in \mathbf{GL}(n, \mathbf{R})$$

telles que $x_{ij} = 0$, pour $i > j$, est un groupe simplement connexe résoluble. Son groupe dérivé est le *groupe trigonalstrict supérieur*, qui est formé des matrices de $\mathrm{T}(n, \mathbf{R})$ telles que $x_{ii} = 1$ pour tout i. Tout groupe de Lie résoluble simplement connexe est isomorphe à un sous-groupe d'un groupe trigonal $\mathrm{T}(n, \mathbf{R})$. On notera qu'un groupe résoluble de dimension $\geqslant 2$ peut avoir son centre réduit à l'élément neutre, par exemple le groupe des matrices réelles :

$$\begin{pmatrix} 1 & y \\ 0 & x \end{pmatrix}$$

où $x > 0$.

Dans un groupe de Lie connexe G, il existe un *plus grand* sous-groupe résoluble connexe R distingué dans G, appelé le *radical* de G ; il est fermé dans G. Lorsque R est réduit à l'élément neutre, on dit que le groupe G est *semi-simple*. Pour un groupe de Lie connexe G, de radical R, $S = G/R$ est semi-simple. Pour qu'un groupe connexe soit semi-simple, il faut et il suffit que son revêtement universel le soit. Un groupe de Lie simplement connexe G est produit semi-direct de son radical R et d'un groupe semi-simple L isomorphe à G/R. Le centre d'un groupe semi-simple est discret.

2. Groupes de Lie compacts et groupes semi-simples

Soit G un groupe de Lie connexe ; il existe alors dans G un sous-groupe compact *maximal* K et un nombre fini de sous-groupes fermés $H_1, \ldots, H_p$ isomorphes à $\mathbf{R}$, tels que l'application :

$$(k, x_1, ..., x_p) \mapsto kx_1 ... x_p$$

du produit :

$$K \times H_1 \times \ldots \times H_p$$

soit un isomorphisme de la variété sous-

jacente à ce produit sur la variété sous-jacente à G ; en outre, pour tout sous-groupe compact K, de G, il existe $s \in G$ tel que :

$$s K_1 s^{-1} \subset K,$$

et en particulier deux sous-groupes compacts maximaux sont conjugués. Les propriétés topologiques de G (par exemple ses groupes d'homotopie ou d'homologie) sont donc connues lorsqu'on connaît les propriétés correspondantes de K.

On peut citer deux exemples : dans $SL(n, R)$, le groupe $SO(n, R)$ est un sous-groupe compact maximal ; dans $GL(n, C)$, le groupe U(n, C), aussi noté U(n), est un sous-groupe compact maximal.

Le revêtement universel d'un groupe de Lie compact K est de la forme K' X $\mathbf{R}^n$, où K' est compact, semi-simple et simplement connexe. Tout groupe compact semi-simple et simplement connexe est produit direct de sous-groupes compacts simplement connexes et *simples* (c'est-à-dire n'ayant pas de sous-groupe fermé distingué distinct d'eux-mêmes et de dimension strictement positive) ; leurs centres sont finis, et les sous-groupes distingués fermés d'un groupe simple sont contenus dans le centre.

Les groupes simples compacts simplement connexes sont explicitement connus (classification de Killing-É. Cartan) : il y a d'abord quatre séries infinies de *groupes classiques* (tabl. 1).

Les groupes de types B, C peuvent être définis pour $m \geqslant 1$ et ceux du type D pour $m \geqslant 2$, mais on n'obtient pas de groupes essentiellement nouveaux, car on a les isomorphismes $A_1 \simeq B_1 \simeq C_1$, $B_2 \simeq C_2$ et $A_3 \simeq D_3$, et le groupe de type D_2 est isomorphe au produit de deux groupes de type A_1. Il faut enfin préciser que le groupe unitaire $U(m, H)$ sur le corps des quaternions H se rapporte à une forme unitaire positive non dégénérée.

Il existe en outre cinq *groupes exceptionnels*, notés :

G_2	14	1
F_4	52	1
E_6	78	3
E_7	133	2
E_8	248	1

(la seconde colonne indique la dimension, et la troisième, l'ordre du centre).

On verra plus loin (chap. 2, 3 et 4) d'autres précisions sur ces groupes. Mentionnons ici que l'algèbre de cohomologie des groupes classiques est entièrement déterminée sur l'anneau des entiers ou sur un corps premier ; on connaît aussi les groupes d'homotopie :

$$\pi_k(U(n, C))$$

pour $k \leqslant 2n + 2$; en particulier :

$$\pi_k(U(n, C)) = Z$$

pour k impair $< 2n$;

$$\pi_k(U(n, C)) = 0$$

type	dimension	ordre du centre	cas
$A_m = SU(m+1, C)$	$m(m+2)$	$m+1$	$m \geqslant 1$
$B_m = Spin(2m+1, R)$	$m(2m+1)$	2	$m \geqslant 2$
$C_m = U(m, H)$	$m(2m+1)$	2	$m \geqslant 3$
$D_m = Spin(2m, R)$	$m(2m-1)$	4	$m \geqslant 4$

tabl. 1 *Groupes classiques simples. compacts et simplement connexes*

pour k pair $< 2n$, et

$$\pi_{2n}(\mathbf{U}(n, \mathbf{C}))$$

est cyclique d'ordre n ! ; on obtient des résultats analogues pour les groupes d'homotopie de $\mathbf{SO}(n, \mathbf{R})$ (théorèmes de Bott).

Les groupes semi-simples *complexes* correspondent biunivoquement aux groupes semi-simples *compacts*, tout groupe semi-simple compact K étant sous-groupe compact maximal d'un groupe semi-simple complexe G, déterminé à isomorphie près, de dimension complexe égale à la dimension de K et dont le centre est celui de K (cf. chap. 6 et 7). Pour les groupes compacts classiques, les groupes simples complexes simplement connexes correspondants sont les suivants :

$$\mathbf{A}_m = \mathbf{SL}(m+1, \mathbf{C});$$
$$\mathbf{B}_m = \mathbf{Spin}(2m+1, \mathbf{C})$$

(revêtement universel de $\mathbf{SO}(2m+1, \mathbf{C})$) ;

$$\mathbf{C}_m = \mathbf{Sp}(2m, \mathbf{C});$$
$$\mathbf{D}_m = \mathbf{Spin}(2m, \mathbf{C})$$

(revêtement universel de $\mathbf{SO}(2m, \mathbf{C})$).

La situation est plus compliquée pour les groupes semi-simples *réels* non compacts (et non sous-jacents à un groupe semi-simple complexe) ; ils peuvent avoir un centre infini (discret) et ne contenir aucun sous-groupe compact distinct de $\{e\}$ (par exemple le revêtement universel de $\mathbf{SL}(2, \mathbf{R})$). On se limitera ici aux groupes semi-simples réels dont le centre est *fini* (le quotient d'un groupe semi-simple par son centre, cf. chap. 5, a toujours un centre réduit a e). Un tel groupe G_c de dimension n est sous-groupe fermé d'un groupe semi-simple complexe G_c, bien déterminé à isomorphic près (le « complexifié » de G, cf. chap. 7), de dimension complexe n ; mais à un même groupe semi-simple complexe G' correspondent *plusieurs* groupes semi-simples réels non isomorphes, dont G' est le complexifié ; on dit que ces groupes sont les « formes réelles » de G' ; une d'entre elles est toujours le groupe compact correspondant à G'. Cependant, si, par exemple, on considère, pour un entier $m \geqslant 2$ donné et pour chaque p tel que $1 \leqslant p \leqslant m$, le groupe orthogonal réel unimodulaire :

$$\mathbf{SO}(p, 2m-p)$$

correspondant à une forme quadratique de signature $(p, 2m-p)$, tous ces groupes sont des formes réelles, deux à deux non isomorphes, du groupe semi-simple complexe $\mathbf{SO}(2m, \mathbf{C})$. Toutes les formes réelles des groupes simples complexes ont été déterminées par É. Cartan.

Un groupe semi-simple réel connexe non compact G admet toujours une *décomposition d'Iwasawa* $G = KAN$, où K est un groupe compact maximal de G, A un groupe commutatif fermé dans G, isomorphe à un $\mathbf{R}^n$, et N un groupe résoluble simplement connexe (donc ayant une variété sous-jacente isomorphe à un $\mathbf{R}^q$) fermé dans G ; le centre de G est contenu dans KA ; en outre, l'application :

$$(k, a, n) \mapsto kan$$

est un isomorphisme de la variété sous-jacente à :

$$\mathbf{K} \times \mathbf{A} \times \mathbf{N}$$

sur la variété sous-jacente à G. Par exemple, si $G = \mathbf{SL}(n, \mathbf{R})$, on peut prendre pour K le groupe orthogonal $\mathbf{SO}(n, \mathbf{R})$, A est le groupe des matrices diagonales de déterminant 1, et N le groupe trigonal strict. Si :

$$\mathbf{G} = \mathbf{SO}(p, 2m-p),$$

on peut prendre pour K le produit :

$$SO(p) \times SO(2m - p).$$

Revenons aux groupes de Lie *compacts*. Un tel groupe K contient des sous-groupes compacts connexes *commutatifs*, donc isomorphes à des *tores* T^k. Un tore *maximal* T dans K est son propre centralisateur (donc contient le centre de K) et, pour tout autre tore T' ⊂ K, il existe un $s \in K$ tel que $sT's^{-1} \subset T$; en particulier, deux tores maximaux sont toujours conjugués dans K. En outre, tout élément de K appartient à au moins un tore maximal. Ainsi, dans le groupe unitaire U(n, C), un groupe compact maximal est formé des matrices diagonales :

$$\text{diag}\,(e^{i\theta_1}, e^{i\theta_2}, ..., e^{i\theta_n})\,;$$

le fait que la réunion des tores maximaux est U(n, C) équivaut ici à la classique réduction d'une matrice unitaire à la forme diagonale par similitude.

La dimension m d'un tore maximal de K est appelée le *rang* de K ; lorsque K est simple, le rang est l'indice m affixé à la lettre A, B..., G dans la classification de Killing-Cartan. Le *normalisateur* N(T) d'un tore maximal T d'un groupe semi-simple compact K joue un rôle important : le groupe quotient W = N(T)/T est appelé *groupe de Weyl* du groupe K (cf. chap. 6).

3. Actions des groupes de Lie

Les groupes de Lie ont d'abord été étudiés en tant que groupes de transformations de certains espaces, plutôt que pour eux-mêmes ; et, dans la théorie moderne, les diverses façons dont un groupe de Lie peut être considéré comme groupe de transformations jouent encore un grand rôle. Les *actions ou opérations* d'un groupe de Lie se définissent comme pour les groupes quelconques (cf. la partie D ci-dessus Représentation linéaire des groupes), mais on n'envisage d'ordinaire que des actions d'un groupe de Lie G sur une variété analytique X, et on exige que l'application $(s, x) \mapsto s\,x$ de G × X dans X soit analytique. Pour tout $s \in G$, l'application $x \mapsto s\,x$ est alors un isomorphisme de la variété X sur elle-même ; pour tout $x \in X$, l'ensemble S_x des $s \in G$ tels que $s\,x = x$ est un sous-groupe fermé de G appelé *stabilisateur* de x. *L'orbite* $G\,x$ de x est l'ensemble des $s\,x$ pour $s \in G$; les orbites sont les classes d'équivalence d'une relation d'équivalence R dans G ; elles ne sont pas nécessairement fermées dans X et peuvent être en fait des ensembles très compliqués. Leur étude générale n'a guère été poussée que pour G = R ou G = **Z**.

L'ensemble X/G des orbites ne peut en général être muni d'une structure de variété analytique telle que l'application canonique $\pi : X \rightarrow X/G$ (qui fait correspondre à un point son orbite) soit une submersion ; pour qu'il en soit ainsi, il faut et il suffit que l'ensemble $\Gamma_R \subset X \times X$ des couples (x, y) appartenant à une même orbite soit une sous-variété fermée de X × X ; toute orbite est alors une sous-variété fermée de X.

Un cas ou la variété des orbites existe toujours est celui où G est un sous-groupe fermé d'un groupe de Lie H, le groupe G opérant dans H par translation à droite $(s, x) \mapsto xs$ avec $s \in G$, $x \in H$, de sorte que les orbites sont les classes à gauche xG dans H. La variété des orbites H/G est alors appelée *l'espace homogène* des classes à gauche suivant G ; le groupe de Lie H opère à gauche sur H/G par $(z, xG) \mapsto zxG$. Lorsqu'un groupe de Lie G opère sur une variété X de sorte que la variété des orbites X/G soit définie, l'orbite d'un point

x est canoniquement isomorphe à l'espace homogène G/S_x.

Le fait pour une variété analytique de X de pouvoir être considérée comme espace homogène H/G d'un groupe de Lie H implique l'existence sur X d'une « géométrie » où se reflètent les propriétés des groupes H et G : c'est l'idée directrice exprimée d'abord par F. Klein dans son programme d'Erlangen, et la géométrie euclidienne classique n'apparaît plus ainsi que comme un exemple particulier des « géométries » associées aux groupes de Lie ; les plus intéressantes correspondent au cas où H est un groupe *simple* (cf. chap. 2) et on a par exemple développé ainsi les « géométries de Tits-Freudenthal » correspondant aux cinq groupes exceptionnels.

Les espaces homogènes G/H les plus importants dans toutes sortes d'applications sont les espaces *riemanniens symétriques irréductibles,* découverts et entièrement énumérés par É. Cartan au cours de recherches de géométrie riemannienne : ce sont les espaces de la forme G/K, où G est un groupe simple réel de centre fini et K un sous-groupe *compact* de G, obtenu comme l'ensemble des $x \in G$ tels que $\sigma(x) = x$ où σ est une *involution* analytique de G (cf. chap. 7). Lorsque G est non compact, K est nécessairement un sous-groupe compact maximal de G, et G/K est difféomorphe à un espace $\mathbf{R}^n$. Si l'on prend $G = \mathbf{SL}(n, \mathbf{R})$, par exemple, $K = \mathbf{SO}(n, \mathbf{R})$ est l'ensemble des matrices invariantes par l'involution $U \mapsto \sigma(U) = {}^tU^{-1}$ (contragédiente de U) ; pour $n = 2$, l'espace symétrique G/K s'identifie canoniquement avec le *demi-plan de Poincaré* formé des nombres complexes de parties imaginaires strictement positives, où la matrice :

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

de G opère par :

$$z \mapsto \frac{az + b}{cz + d},$$

avec a, b, c, d réels et $ad - bc = 1$. Dans beaucoup de cas (entièrement déterminés par É. Cartan), les espaces symétriques G/K s'identifient ainsi à des ouverts d'espaces *complexes* $\mathbf{C}^n$, *où* G opère par transformations *holomorphes,* et ces espaces jouent un rôle important dans la théorie des fonctions de plusieurs variables complexes. Plus récemment, on a pu déterminer également les ouverts bornés de $\mathbf{C}^n$ qui sont des espaces homogènes G/H (non nécessairement symétriques) où G opère par transformations holomorphes. Les *sphères* et les *espaces projectifs* sont aussi des espaces riemanniens symétriques irréductibles.

4. Représentations linéaires de dimension finie des groupes de lie

Les définitions sont données à l'article précédent, qui traite de la représentation linéaire des groupes ; on se bornera aux représentations linéaires dans des espaces vectoriels V (de dimension *finie* dans ce chapitre) sur le corps C des nombres *complexes* ; en outre, les représentations linéaires $\rho : G \to GL(V)$ d'un groupe de Lie que l'on considère sont supposées *analytiques* (réelles).

Lorsque le groupe de Lie G est connexe et *résoluble,* toute représentation irréductible de G est de dimension 1, autrement dit de la forme $s \mapsto \chi(s)$, où χ est un caractère (abélien) du groupe commutatif G/D(G) ; une représentation quelconque

de G s'écrit toujours sous la forme triangulaire :

$$s \mapsto \begin{pmatrix} \chi_1(s) & \alpha_{12}(s) & \dots & \alpha_{1p}(s) \\ 0 & \chi_2(s) & \dots & \alpha_{2p}(s) \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \chi_p(s) \end{pmatrix}$$

L'exemple de la représentation linéaire :

$$t \mapsto \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

de G = R montre qu'une représentation linéaire d'un groupe commutatif n'est pas nécessairement complètement réductible.

En revanche, toute représentation linéaire d'un groupe de Lie G *compact* ou *réductif* (c'est-à-dire dont le revêtement universel est produit d'un groupe semi-simple et d'un $\mathbf{R}^n$) est *complètement réductible* (théorème de H. Weyl) ; pour les groupes compacts, c'est même vrai sans supposer que G est un groupe de Lie (cf. art. précédent). Tout revient donc à déterminer, dans ces cas, les représentations *irréductibles* ; cette détermination a été complètement effectuée par É. Cartan au moyen de techniques qui seront esquissées dans le chapitre 6.

La théorie des représentations linéaires des groupes semi-simples généralise la théorie classique des *invariants* en géométrie projective. 11 s'agissait uniquement, dans cette théorie, des représentations des groupes classiques, et surtout de SL(V). Ce groupe opère en effet naturellement dans **toute** puissance tensorielle $V^{\otimes n}$, et dans le sous-espace des tenseurs symétriques d'ordre n. Ce dernier s'identifie à l'espace vectoriel F,, des polynômes homogènes de degré n à p variables (si p = dim V) ; un élément $s \in$ SL(V) opère en transformant un tel polynôme $P_n(x_1, \dots . x_p)$ en le polynôme :

$$s P_n(x_1, \dots . x_p) = P_n(s^{-1}. x_1, \dots . s^{-1}. x_p),$$

en posant $s^{-1} x_k = a_{k1}x_1 + \dots + a_{kp}x_p$ si s^{-1} est la matrice (cc,). Un invariant dans F_n est un polynôme tel que $s P_n = P_n$ pour tout $s \in$ SL(V) ; cela signifie que P_n engendre dans F_n un sous-espace stable de *dimension* 1, et, si l'on sait décomposer toute représentation linéaire en représentations irréductibles, on pourra obtenir tous les invariants. D. Hilbert avait démontré (pour SL(V)) qu'il y a un nombre *fini* de polynômes homogènes invariants I_1, I_2, ..., I_r tel que tout autre invariant soit de la forme $Q(I_1, I_2, \dots, I_r)$, où Q est un polynôme. Ce théorème s'étend à *tous les groupes semi-simples* (mais non à tous les groupes de Lie).

Une représentation linéaire :

$$\rho : G \rightarrow GL(V)$$

est dite *fidèle* si elle est injective. On peut prouver que, pour tout groupe de Lie connexe G, il existe un groupe connexe qui a même revêtement universel que G et qui est isomorphe à un sous-groupe d'un groupe linéaire GL(V) ; mais le revêtement universel de G n'a pas toujours cette propriété (par exemple pour G = SL(2, R)). Toutefois, tout groupe compact et tout groupe semi-simple ***complexe*** est isomorphe à un sous-groupe d'un groupe linéaire.

5. Algèbres de Lie

L'outil essentiel dans la démonstration des remarquables résultats qui précèdent est la méthode infinitésimale, inaugurée par S. Lie (1842-1899), qui a pour effet de ramener l'étude des groupes de Lie à

l'étude de ce qu'on appelle leurs *algèbres de Lie*. L'idée est d'étudier les conditions qu'impose l'associativité de la loi d'un groupe G aux séries qui l'expriment dans un voisinage V de e. On suppose choisi un système de coordonnées locales qui s'annulent en e, de sorte que V est identifié à un voisinage de l'origine dans $\mathbf{R}^n$. Soit W un voisinage de 0 tel que $W^2 \subset V$, et $x = (x_1, ..., x_n)$, $y = (y_1, ..., y_n)$ deux points de W; leur produit $z = xy \in V$ étant fonction analytique de x, y par hypothèse, les coordonnées $z_1, ..., z_n$ de z s'expriment par des séries convergentes pour $|x_j| < \rho$, $|y_j| < \rho$, $1 \leqslant j \leqslant n$:

$$(1) \quad z_j = \varphi_j(x,y) = x_j + y_j + \sum_{|\alpha| \geqslant 1, |\beta| \geqslant 1} b^{(j)}_{\alpha\beta} x^\alpha y^\beta,$$

$1 \leqslant j \leqslant n$, où l'on a employé la notion des multi-indices :

$$\alpha = (\alpha_1, ..., \alpha_n),$$
$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} ... x_n^{\alpha_n},$$
$$|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_n$$

(cf. CALCUL INFINITÉSIMAL Calcul à plusieurs variables). Considérons alors une fonction analytique :

$$f(x) = f(x_1, ..., x_n),$$

donc développable en série convergente au voisinage de 0 ; si l'on substitue à chaque x_j la série $z_j = \varphi_j(x, y)$ donnée par (1), on obtient une série en les x_j et y_j, et, en groupant les monômes en $x^\beta y^\alpha$ pour un même α, on obtient ce qu'on peut appeler la *formule de Taylor* dans le groupe G au voisinage de e :

$$(2) \quad f(xy) = \sum_\alpha (Z_\alpha f(x)) y^\alpha,$$

où on vérifie aisément que :

$$Z_\alpha f(x) = \sum_{\beta \leqslant \alpha} a_{\alpha\beta}(x) D^\beta f(x)$$

(combinaison d'un nombre fini de dérivées partielles de f, à coefficients $a_{\alpha\beta}$ analytiques au voisinage de 0). Les applications $f \mapsto Z_\alpha f$ sont donc des *opérateurs différentiels* sur les fonctions analytiques dans G ; en outre, ils ont la propriété fondamentale *d'invariance à gauche* par le groupe. De façon précise, pour tous $s \in G$ assez petit, posons $f_s(x) = f(sx)$. Un opérateur différentiel Z est dit invariant à gauche si $Z(f_s) = (Z(f))_s$, pour s assez petit ; pour les Z_α, cela résulte de leur définition (2) et de l'associativité, qui donne $f_s(xy) = f((sx)y)$. Il est clair que l'ensemble $\mathfrak{G}$ des opérateurs différentiels invariants à gauche est une *algèbre associative* sur R, dont on voit aisément que les Z_α forment une *base* sur R (Z_0 est pris égal à l'identité). En fait, la table de multiplication de la base (Z_α) se détermine explicitement à l'aide des séries (1). On pose en effet, pour tout multi-indice $\gamma = (\gamma_1, ..., \gamma_n)$:

$$(3) \quad z^\gamma = (\varphi_1(x,y))^{\gamma_1} ... (\varphi_n(x,y))^{\gamma_n} = \sum_{\alpha,\beta} c_{\alpha\beta\gamma} x^\alpha y^\beta,$$

de sorte que $b^{(j)}_{\alpha\beta} = c_{\alpha\beta\varepsilon_j}$, où ε_j est le multi-indice (δ_{ij}), avec $1 \leqslant i \leqslant n$; de plus, on vérifie aussitôt que $c_{\alpha\beta\gamma} = 0$ pour $\alpha + \beta < \gamma$ et que les seuls $c_{\alpha\beta\gamma}$ non nuls tels que $|\gamma| = |\alpha| + |\beta|$ sont ceux pour lesquels $\gamma = \alpha + \beta$, qui ont pour valeur :

$$c_{\alpha,\beta,\alpha+\beta} = \frac{(\alpha + \beta)!}{\alpha! \beta!},$$

avec $\alpha! = \alpha_1! ... \alpha_n!$. Pour tout s assez petit, on peut écrire :

$$f(sxy) = \sum_\beta (Z_\beta f(sx)) y^\beta,$$

et, en vertu de la formule (2) appliquée en remplaçant f par $Z_\beta f$:

$$(4) \qquad f(sxy) = \sum_{\alpha,\beta} ((Z_\alpha Z_\beta f)(s)) x^\alpha y^\beta .$$

Mais, d'autre part, on a aussi, par (2) :

$$f(sxy) = \sum_{\gamma} (Z_\gamma f(s))(xy)^\gamma$$

$$= \sum_{\gamma} (Z_\gamma f(s)) \Big(\sum_{\alpha,\beta} c_{\alpha\beta\gamma} x^\alpha y^\beta \Big);$$

d'où, en comparant à (4), les formules :

$$(5) \qquad Z_\alpha Z_\beta = \sum_{\gamma} c_{\alpha\beta\gamma} Z_\gamma ,$$

qui donnent la table de multiplication. La comparaison des formules (3) et (5) montre que la structure d'algèbre de $\mathfrak{G}$ et la structure de groupe de G (si G est simplement connexe) *se déterminent mutuellement* sans ambiguïté.

En particulier, en posant :

$$Z_{\varepsilon_i} = X_i, \quad b^{(k)}_{\varepsilon_i \varepsilon_j} = b_{ijk},$$

on tire de (5), en prenant $\alpha = \varepsilon_i$, $\beta = \varepsilon_j$:

$$(6) \qquad X_i X_j = Z_{\varepsilon_i + \varepsilon_j} + \sum_{k=1}^{n} b_{ijk} X_k ,$$

$1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant n$; et, en échangeant i et j :

$$(7) \quad [X_i, X_j] = X_i X_j - X_j X_i = \sum_{k=1}^{n} (b_{ijk} - b_{jik}) X_k ,$$

$1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant n$.

Le sous-espace vectoriel $\mathfrak{g}$ de $\mathfrak{G}$, de dimension n, admettant les X_i pour base, est l'ensemble des opérateurs invariants à gauche d'ordre 1 ; les formules (7) montrent que $\mathfrak{g}$ est stable pour l'opération :

$$(X, Y) \mapsto [X, Y] = XY - YX,$$

qui vérifie les deux identités :

$$(8) \qquad [X, Y] = -[Y, X],$$

$$(9) \qquad [X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$$

(identité de Jacobi). Un espace vectoriel sur un corps K dans lequel est défini une loi de composition :

$$(u, v) \mapsto [u, v]$$

vérifiant ces deux identités et bilinéaire est appelé *algèbre de Lie* sur K. Un homomorphisme $f: \mathfrak{g} \to \mathfrak{g}'$ d'algèbres de Lie sur le même corps est par définition une application K-linéaire telle que :

$$f([u, v]) = [f(u), f(v)].$$

On a donc associé canoniquement à tout groupe de Lie G une *algèbre de Lie* $\mathfrak{g}$ sur R, dite algèbre de Lie de ce groupe et notée Lie (G). Il est très facile, en partant des formules (4) et (5) et des propriétés des $c_{\alpha\beta\gamma}$, de voir que les X_i engendrent l'algèbre associative $\mathfrak{G}$; de façon précise, les monômes :

$$X_\alpha = X_1^{\alpha_1} X_2^{\alpha_2} \dots X_n^{\alpha_n}$$

forment une base de l'espace vectoriel $\mathfrak{G}$ (l'ordre des facteurs dans les X_α est bien entendu essentiel). De plus $\mathfrak{G}$ est entièrement déterminée à isomorphie près lorsqu'on connaît $\mathfrak{g}$, car elle est caractérisée par la propriété « universelle » suivante : Pour toute application linéaire f de $\mathfrak{g}$ dans une R-algèbre associative A telle que :

$$f([X, Y]) = f(X)f(Y) - f(Y)f(X),$$

il existe un homomorphisme d'algèbres F de $\mathfrak{G}$ dans A et un seul qui prolonge f.

Lorsque G est un groupe de Lie complexe, son algèbre de Lie $\mathfrak{g}$ est une algèbre de Lie sur le corps C ; quand on la considère comme algèbre de Lie sur R, elle

est l'algèbre de Lie du groupe réel sous-jacent à G.

Voici quelques exemples. Si G = $\mathbf{R}^n$, on a :

$$Z_\alpha = \frac{1}{\alpha!} D^\alpha$$

et (2) est la formule de Taylor usuelle ; l'algèbre associative $\mathfrak{G}$ s'identifie à l'algèbre des polynômes en les D,, $1 \leqslant j \leqslant n$; l'algèbre de Lie correspondante est *commutative,* c'est-à-dire que [X, Y] = 0 quels que soient X, Y dans g. L'algèbre de Lie du groupe des matrices :

$$\begin{pmatrix} 1 & Y \\ 0 & x \end{pmatrix}$$

a une base de deux éléments X, Y vérifiant la table de multiplication [X, Y] = -Y. L'algèbre de Lie $\mathfrak{gl}(n$, R) (resp. $\mathfrak{gl}(n$, C)) du groupe linéaire $\mathbf{GL}(n$, R) (resp. $\mathbf{GL}(n$, C)) s'identifie canoniquement à l'espace des matrices carrées réelles (resp. complexes) d'ordre n où le crochet est l'application $(X, Y) \mapsto XY - YX$; l'algèbre de Lie de $\mathbf{SL}(n$, R) est la sous-algèbre de Lie $\mathfrak{sl}(n$, R) de $\mathfrak{gl}(n$, R) formée des matrices de trace 0. En particulier, $\mathfrak{sl}(2$, R) (ou $\mathfrak{sl}(2$, C)) a pour base les trois matrices :

$$H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad Y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

vérifiant donc la table de multiplication :

(10) $$[H, X] = 2X, \quad [H, Y] = -2Y, \quad [X, Y] = H.$$

On montre que réciproquement, à *toute* algèbre de Lie 5 sur R (resp. C) de dimension finie, correspond un groupe de Lie réel (resp. complexe) simplement connexe et un seul à isomorphie près, dont l'algèbre de Lie est isomorphe à g. Tous les groupes de Lie connexes ayant la même algèbre de Lie ont même revêtement universel (à isomorphie près). Cette correspondance permet d'établir un « dictionnaire » entre les notions fondamentales de la théorie des groupes de Lie et des notions de la théorie des algèbres de Lie, qui relèvent essentiellement de l'algèbre linéaire (pour deux sous-espaces vectoriels a, b d'une algèbre de Lie 5, on note dans ce qui suit [a, b] le sous-espace vectoriel engendré par les [X, Y] pour X $\in$ a, Y $\in$ b) selon le tableau 2.

Pour tout $s \in$ G, il correspond à l'automorphisme intérieur $\mathrm{Int}(s) : x \mapsto sxs^{-1}$ de G l'automorphisme dérivé $(\mathrm{Int}(s))_*$ de $\mathfrak{g}$, noté Ad(s) ; l'application $s \mapsto$ Ad(s) de G dans GL(g) est une représentation linéaire de G dans l'espace vectoriel $\mathfrak{g}$, appelée *représentation adjointe* ; son noyau est le centre Z de G et son image Ad(G) C $\mathbf{GL}(\mathfrak{g})$, isomorphe à G/Z, est appelée le *groupe adjoint* de G. On montre que la représentation linéaire Ad, de $\mathfrak{g}$ dans $\mathfrak{gl}(\mathfrak{g})$ est l'application X $\mapsto$ ad(X), où on pose :

$$\mathrm{ad}(X)(Y) = [X, Y] ;$$

cette application est dite *représentation adjointe* de 5 ; son noyau est le centre $\mathfrak{z}$ de 5 et, pour tout XE $\mathfrak{g}$, ad(X) est une *dérivation* de l'algèbre 5, c'est-à-dire :

$$\mathrm{ad}(X)[Y, Z] = [\mathrm{ad}(X)(Y), Z] + [Y, \mathrm{ad}(X)(Y)],$$

ce qui n'est autre que l'identité de Jacobi.

Pour tout X $\in \mathfrak{g}$, il existe un homomorphisme et un seul du groupe additif R dans G, dont l'homomorphisme dérivé soit $t \mapsto tX$; on note cet homomorphisme $t \mapsto \exp(tX)$ et l'image de R par cet homomorphisme est appelée le *sous-groupe à un paramètre* de G correspondant à X. Un tel sous-groupe n'est pas nécessairement fermé dans G ; ainsi, par exemple, si G $= \mathbf{T}^2$ et si φ : R $\mapsto$ T est l'homomorphisme canonique, alors $t \mapsto$ (<p(t), $\varphi(\theta t)$), où θ est un nombre irrationnel,

groupe de Lie simplement connexe G	algèbre de Lie $\mathfrak{g}$
homomorphisme $f : G \to G'$ de groupes de Lie simplement connexes	homomorphisme dérivé $f_* : \mathfrak{g} \to \mathfrak{g}'$ d'algèbres de Lie défini par : $$(f_*(X))(\varphi) = X(\varphi \circ f)$$ pour tout opérateur $X \in \mathfrak{g}$ et toute fonction différentiable φ sur G'
image de f	Image de f_* (sous-algèbre de $\mathfrak{g}'$)
composante neutre du noyau de f (sous-groupe distingué fermé)	noyau de f_* (idéal de $\mathfrak{g}$, c'est-à-dire sous-espace $\mathfrak{a} \subset \mathfrak{g}$ tel que : $$[X, Y] \in \mathfrak{a} \quad \text{pour } X \in \mathfrak{g}, Y \in \mathfrak{a}$$
quotient G/N de G par un sous-groupe distingué fermé N	quotient $\mathfrak{g}/\mathfrak{n}$ de l'algèbre de Lie de G par celle de la composante neutre de N
composante neutre du centre	centre (idéal $\mathfrak{z}$ des X tels que [X, Y] = 0 pour tout $Y \in \mathfrak{g}$)
groupe commutatif	algèbre de Lie commutative
sous-groupe (G, H) pour un sous-groupe distingué fermé connexe H de G (le groupe (G, H) est engendré par les $ghg^{-1}h^{-1}$ pour $g \in G$, $h \in H$)	idéal $[\mathfrak{g}, \mathfrak{h}]$ formé des combinaisons linéaires des [X, Y] pour $X \in \mathfrak{g}$, $Y \in \mathfrak{h}$
groupe dérivé (ou groupe des commutateurs)	algèbre dérivée $[\mathfrak{g}, \mathfrak{g}]$ de $\mathfrak{g}$
produit direct G × G'	produit direct $\mathfrak{g} \times \mathfrak{g}'$, avec : $$[(X, X'), (Y, Y')] = ([X, Y], [X', Y'])$$
produit semi-direct	produit semi-direct d'un idéal $\mathfrak{a}$ et d'une sous-algèbre $\mathfrak{b}$, avec : $$\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b}$$
représentation linéaire de G dans V : $$\rho : G \to GL(V)$$	représentation linéaire de $\mathfrak{g}$ dans V: $$\rho_* : \mathfrak{g} \to \mathfrak{gl}(V)$$

tabl. 2 • *Correspondance entre la théorie des groupes de Lie et la théorie des algèbres de Lie*

définit un sous-groupe à un paramètre partout dense dans $\mathbf{T}^2$. On a :

$$\exp((t + t')X) = \exp(tX)\exp(t'X),$$
$$\exp(s(tX)) = \exp((st)X) ;$$

par contre, si X et Y sont tels que $[X, Y] \neq 0$, on a en général :

$$\exp(t(X + Y)) \neq \exp(tX)\exp(tY).$$

On montre qu'il existe un voisinage assez petit V de 0 dans $\mathfrak{g}$ tel que l'*application exponentielle* $X \mapsto \exp(X)$ de $\mathfrak{g}$ dans G, restreinte à V, soit un isomorphisme analytique de V sur un voisinage ouvert de e ; mais en général, l'application $X \mapsto \exp(X)$ n'est ni injective ni surjective dans $\mathfrak{g}$. Elle est toutefois surjective lorsque G est compact, et bijective lorsque G est résoluble et simplement connexe. On prouve, en outre, que :

$$\text{(11)} \qquad \mathrm{Ad}(\exp(X)) = \exp(\mathrm{ad}(X))$$

pour $X \in \mathfrak{g}$, l'exponentielle du second membre étant la série usuelle :

$$\exp(A) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n$$

dans l'algèbre des matrices.

6. Algèbres de Lie semi-simples

La notion d'algèbre de Lie *résoluble* (resp. *nilpotente*) se définit comme pour les groupes, en remplaçant les groupes D'(G) (resp. $C^r(G)$) par les idéaux formés de la façon correspondante dans l'algèbre de Lie $\mathfrak{g}$. Si G est un groupe de Lie simplement connexe, R son radical, le plus grand idéal résoluble $\mathfrak{r}$ de l'algèbre de Lie $\mathfrak{g}$ de G est l'algèbre de Lie de R, et on l'appelle le *radical* de g. Une algèbre de Lie $\mathfrak{g}$ est dite *semi-simple* si son radical est réduit à $\{0\}$ (ou, ce qui revient au même, si elle ne contient pas d'idéal commutatif non réduit à $\{0\}$. Un groupe de Lie connexe est semi-simple si et seulement si son algèbre de Lie est semi-simple.

On définit d'autre part sur toute algèbre de Lie réelle (resp. complexe) $\mathfrak{g}$ une *forme bilinéaire symétrique* réelle (resp. complexe) diteforme *de Killing*, par la formule :

$$(X|Y) = \mathrm{Tr}(\mathrm{ad}(X)\mathrm{ad}(Y)).$$

Cette forme est étroitement liée à la structure de $\mathfrak{g}$ par les trois *critères de Cartan* :

Pour que $\mathfrak{g}$ soit résoluble, il faut et il suffit que $(X|Y) = 0$ pour $X \in \mathfrak{g}$ et $Y \in [\mathfrak{g}, \mathfrak{g}]$.
Pour que $\mathfrak{g}$ soit semi-simple, il faut et il suffit que la forme de Killing soit non dégénérée.

Pour qu'une algèbre de Lie réelle $\mathfrak{g}$ soit l'algèbre de Lie d'un groupe compact, il faut et il suffit que $(X|X) \leqslant 0$ dans g.

On peut parvenir à la détermination de la structure d'un groupe *compact* semi-simple G, en analysant sa représentation adjointe. Il est commode de commencer par étendre canoniquement chaque endomorphisme Ad(s) (pour $s \in G$) de l'algèbre de Lie $\mathfrak{g}$ à un endomorphisme de sa complexifiée $\mathfrak{g}_c = \mathfrak{g} \otimes_{\mathbf{R}} \mathbf{C}$, de sorte qu'on peut considérer G comme opérant par s-Ad(s) soit sur $\mathfrak{g}$, soit sur $\mathfrak{g}_c$. L'idée fondamentale est de restreindre la représentation adjointe *à* un *tore maximal* T de G ; comme T est compact et commutatif et que la forme de Killing est invariante par tout automorphisme de $\mathfrak{g}$ et négative non dégénérée, cette représentation est complètement réductible, donc $\mathfrak{g}$ se décompose en somme directe de sous-espaces E_k, deux à deux orthogonaux pour $(X|Y)$, de dimension 1 ou 2 sur R, et stables par Ad(s), $s \in T$; mais le cas $\dim(E_k) = 1$ est à exclure, car le groupe à un paramètre engendré par un élément de E_k commuterait alors avec T, contrairement à l'hypothèse que T est maximal. Alors :

$$E_k \otimes_{\mathbf{R}} \mathbf{C}$$

est somme directe de deux sous-espaces E'_k et E''_k de dimension 1 sur C, dans lesquels on a :

$$\mathrm{Ad}(s)\, X' = \chi_k(s) X',$$
$$\mathrm{Ad}(s).\, X'' = \chi_k(s) X''$$

respectivement, *où* χ_k est un *caractère* de T ; en vertu de (1 1), il revient au même de dire que, pour tout

$$H \in \mathfrak{h} = \mathfrak{t} \otimes_{\mathbf{R}} \mathbf{C},$$

$\mathfrak{t} = \mathrm{Lie}(T)$, on a :

(12) $$[H, X'] = \alpha_k(H) X' \quad \text{dans } E'_k$$
$$[H, X''] = -\alpha_k(H) X'' \quad \text{dans } E''_k$$

avec $\chi_k(\exp(H)) = \exp 2\pi i \alpha_k(H)$, où α_k

est une forme linéaire non identiquement nulle sur $\mathfrak{h}$, à valeurs réelles dans t ; on dit que les α_k sont les *racines* de $\mathfrak{g}_c$ relativement à la sous-algèbre commutative maximale $\mathfrak{h}$. L'identité de Jacobi et le fait que $\mathfrak{h}$ est maximale montrent que $[E'_k, E''_k] \subset \mathfrak{h}$; on constate alors que la somme directe :

$$E'_k \oplus E''_k \oplus [E'_k, E''_k]$$

est une sous-algèbre de $\mathfrak{g}_c$ avec une base vérifiant (10), elle est donc *isomorphe à* $\mathfrak{sl}(2, C)$. Une analyse élémentaire des représentations irréductibles de $\mathfrak{sl}(2, C)$ permet d'obtenir les résultats fondamentaux suivants : toutes les racines α_k $(1 \leqslant k \leqslant (n - m)/2$ si $n = \dim G$, $m = \dim T)$ sont distinctes ; on peut donc prendre leur ensemble R comme ensemble d'indices, écrire $\mathfrak{g}_\alpha$ et $\mathfrak{g}_{-\alpha}$ au lieu de E'_k et E''_k, et déterminer dans chaque $\mathfrak{g}_\alpha$ un élément X_α, de sorte que, si l'on pose :

(13) $$H_\alpha = [X_\alpha, X_{-\alpha}] \in t,$$

on obtienne :

(14) $$\alpha(H_\alpha) = 2.$$

En outre, pour deux racines quelconques α, β, le nombre $\beta(H_\alpha)$ est égal *à* $p - q$, *où* p et q sont deux *entiers* positifs ou nuls tels que les entiers k vérifiant $-p \leqslant k \leqslant q$ soient exactement ceux pour lesquels $\beta + k\alpha$ est une racine (on montre qu'on a toujours $p + q \leqslant 3$) ; enfin :

(15) $$\begin{cases} [X_\alpha, X_\beta] = 0 & \text{si } \alpha + \beta \text{ n'est pas une racine,} \\ [X_\alpha, X_\beta] = N_{\alpha\beta} X_{\alpha+\beta} & \text{si } \alpha + \beta \text{ est une racine,} \end{cases}$$

et l'on peut montrer que :

(16) $$N_{-\alpha,-\beta} = -N_{\alpha\beta} \quad \text{et} \quad N_{\alpha\beta} = \pm(p+1),$$

où l'entier p a été défini ci-dessus (base de Weyl-Chevalley) ; les H_α engendrent t et l'on obtient :

(17) $$\alpha(H) = (H \mid H_\alpha) \quad \text{pour } H \in \mathfrak{h}.$$

On prouve, à l'aide de ces relations, que l'application :

(18) $$s_\alpha : \beta \mapsto \beta - \beta(H_\alpha)\alpha$$

est une *permutation involutive* de l'ensemble R des racines ; de plus, si $\alpha \in R$, on a $2\alpha \notin R$. La détermination complète des ensembles finis R contenus dans le dual $\mathfrak{h}^*$ de l'espace vectoriel $\mathfrak{h}$, ne contenant pas 0, engendrant $\mathfrak{h}^*$ et ayant les deux propriétés précédentes (ensembles dénommés *systèmes de racines*) est essentiellement un problème de géométrie élémentaire, qui conduit à la classification de Killing-Cartan (cf. chap. 2).

Ainsi, pour le type A_m, qui correspond à l'algèbre de Lie :

$$\mathfrak{sl}(m+1, C)$$

du groupe unimodulaire, on peut prendre pour $\mathfrak{h}$ l'algèbre de Lie engendrée par les éléments $E_{ii} - E_{jj}$ (E_{hk} est la matrice ayant un seul élément $\neq 0$, situé dans la h-ième ligne et la k-ième colonne, et égal à 1) ; les racines α_{ij} correspondent aux couples (i, j) tels que $i \neq j$, avec $\alpha_{ji} = -\alpha_{ij}$; on a :

$$H_{\alpha_{ij}} = E_{ii} - E_{jj},$$
$$X_{\alpha_{ij}} = E_{ij}, \quad X_{-\alpha_{ij}} = E_{ji},$$
$$\alpha_{ij}(H_{\alpha_{hk}}) = 0,$$

si ni h ni k ne sont égaux à l'un des indices i, j, et

$$\alpha_{ij}(H_{\alpha_{hk}}) = 1$$

si $h = i$ et $j \neq k$, *ou* $h \neq i$ et $j = k$. On peut vérifier sur cet exemple les formules générales données plus haut.

Les racines α appartiennent en fait au dual t* de l'espace vectoriel réel t ; si on

définit sur t* la forme bilinéaire inverse de (U V) sur t, qu'on note $(\xi \ \eta)$, la permutation s_α est la restriction à R de la *réflexion orthogonale* par rapport à l'hyperplan M_α des $\xi \in t^*$ tels que $\xi(H_\alpha) = 0$. Les s_α pour $\alpha \in R$ engendrent un groupe *fini* de transformations orthogonales, canoniquement isomorphe au groupe de Weyl W de G (cf. chap. 2) auquel on l'identifie. Les composantes connexes, dans l'espace t*, de la réunion des hyperplans M_α sont appelées les *chambres* de $\mathfrak{g}$ (relatives à t) ; pour une chambre C, il y a exactement m hyperplans M_α tels que la réunion des $M_\alpha \cap \bar{C}$ constitue la frontière de $\bar{C}$ (donc $\bar{C}$ est un « angle polyèdre » dans l'espace t* à m dimensions) ; on dit que ces hyperplans M_α sont les *murs* de C. Si on choisit une chambre C, et qu'on note $M_i (1 \leqslant i \leqslant m)$ ses murs, pour chaque i une des deux racines opposées orthogonales à M_i est du même côté que C de M_i ; elle est notée α_i et on dit que les α_i (qui forment une base de l'espace vectoriel t*) forment une *buse* du système de racines R. On prouve que toute racine $\alpha \in R$ est combinaison linéaire des α_i à coefficients *entiers de même signe;* les racines de R sont ainsi divisées en deux classes, dites *positives* (resp. *négatives*) pour C si tous les coefficients sont $\geqslant 0$ (resp. $\leqslant 0$). Le groupe de Weyl permute les chambres (donc aussi les bases de R) de façon *simplement transitive.* Donnons un exemple : pour le type A_m, on peut prendre comme base les racines $\alpha_{i,i+1}$ pour $1 \leqslant i \leqslant m$; donc le groupe de Weyl s'identifie au groupe symétrique Σ_{m+1} des permutations de $m + 1$ objets.

Supposons G simplement connexe. On a *vu* que *les représentations linéaires* de G dans un espace vectoriel complexe V correspondent biunivoquement aux représentations linéaires de $\mathfrak{g}$ dans V, et aussi (puisque V est un espace vectoriel complexe) à celles de $\mathfrak{g}_c$ dans V ; ces dernières sont donc complètement réductibles, et il suffit de déterminer les *représentations irréductibles* de $\mathfrak{g}_c$. On utilise la même idée que ci-dessus, savoir la *restriction* à $\mathfrak{h}$ d'une représentation ρ de $\mathfrak{g}_c$ dans V ; on appelle *poids* de la représentation ρ (irréductible ou non) tout élément $\omega \in \mathfrak{h}^*$ pour lequel il existe un vecteur $x \in V$ non nul et pour lequel :

$$\rho(H) . x = \omega(H) x$$

pour tout $H \in \mathfrak{h}$ (donc x est vecteur propre commun à tous les endomorphismes $\rho(H)$) ; l'ensemble V_ω des vecteurs x ayant cette propriété pour un poids ω est un sous-espace vectoriel V_ω et V est *somme directe* des V_ω. Les poids de ρ sont donc en nombre fini ; si l'on suppose maintenant ρ irréductible et si l'on choisit une base (α_i) de R, avec $1 \leqslant i \leqslant m$, on démontre qu'il existe un *unique* poids π de ρ tel que tous les autres poids de ρ soient de la forme :

$$\pi - \sum_{i=1}^{m} q_i \alpha_i,$$

où les q_i sont des *entiers positif* ; on dit que π est le *poids dominant* de la représentation ρ, et on montre que V_π est de dimension 1. Deux représentations irréductibles de $\mathfrak{g}_c$ ayant même poids dominant sont semblables. Pour qu'une forme linéaire $\omega \in \mathfrak{h}^*$ soit poids dominant d'une représentation irréductible de $\mathfrak{g}_c$, il faut et il suffit que l'on ait $\omega(H_{\alpha i}) \geqslant 0$ pour $1 \leqslant i \leqslant m$. Les poids dominants de toutes les représentations irréductibles de $\mathfrak{g}_c$ forment donc un Z-module libre P(R), ayant pour base les poids π_i $(1 \leqslant i \leqslant m)$ tels que :

$$\pi_i(H_{\alpha_i}) = 1, \ \pi_i(H_{\alpha_j}) = 0$$

pour $i \neq j$; les π_i sont appelés les *poids fondamentaux* de $\mathfrak{g}_c$ (pour la base (α_i)) ; on

a évidemment $R \subset P(R)$ et le sous-groupe $Q(R)$ engendré par R est d'indice fini dans $P(R)$; on prouve que le quotient $P(R)/Q(R)$ est isomorphe au *centre* de G (supposé simplement connexe).

Ainsi, pour le type A_m, les poids fondamentaux ω_i $(1 \leqslant i \leqslant m)$ sont donnés par :

$$\omega_i(\mathrm{diag}(t_1, \ldots, t_{m+1})) = t_1 + t_2 + \cdots + t_i,$$

qui est une fonction restreinte à la sous-algèbre $\mathfrak{h}$ des matrices diagonales telles que :

$$\sum_{i=1}^{m+1} t_i = 0.$$

On vérifie que ω_1, est le poids dominant de la représentation identique :

$$\mathfrak{sl}(m+1, \mathbf{C}) \rightarrow \mathfrak{sl}(m+1, \mathbf{C})\ ;$$

on montre que ω_j est le poids dominant de la représentation canonique dans la *puissance extérieure j-ième* de $\mathbf{C}^{m+1}$.

Le *caractère* (cf. art. précédent) de la représentation irréductible du groupe compact semi-simple G de poids dominant ω est donné par la *formule de H. Weyl* :

$$(19)\quad \chi_\omega = \frac{\sum_{w \in W} \varepsilon(w) \exp 2\pi i w(\omega + \sigma)}{\sum_{w \in W} \varepsilon(w) \exp 2\pi i w(\sigma)}$$

où σ est la *demi-somme* des racines positives, w parcourt le groupe de Weyl et $\varepsilon(w)$ est son déterminant (égal à ± 1 suivant que w est produit d'un nombre pair ou impair de réflexions s_α) ; le caractère apparaît comme une fonction définie dans l'algèbre de Lie $\mathfrak{t}$ de T, mais a la même valeur pour tous les éléments $H \in \mathfrak{t}$ tels que $\exp(H) \in T$ ait la même valeur dans T (cela résulte de ce que σ est la somme des poids fondamentaux ω_i) ; en fait, χ_ω est donc définie dans T, et, comme tout élément de G est contenu dans un conjugué de T, χ_ω est bien définie dans G.

En particulier, la *dimension* de l'espace de la représentation irréductible de poids dominant ω s'obtient en prenant la valeur de $\chi_\omega(H)$ pour $H = 0$, et on montre que cette valeur est :

$$(20)\qquad \prod_{\alpha > 0} \frac{(\alpha \mid \omega + \rho)}{(\alpha \mid \rho)}$$

(produit étendu aux racines positives).

7. Algèbres semi-simples complexes et leurs formes réelles

Dans le chapitre 6, en partant de l'algèbre de Lie d'un groupe semi-simple compact, on a obtenu, en la complexifiant, une algèbre de Lie semi-simple complexe. Ce processus admet une réciproque, qui établit une correspondance biunivoque entre groupes connexes semi-simples *complexes* et groupes connexes semi-simples *compacts*.

L'unique méthode connue pour établir ce fait est due à Killing et É. Cartan, et est fort longue : on commence par démontrer, dans une algèbre semi-simple *complexe* $\mathfrak{g}$ de dimension n sur C, l'existence d'une sous-algèbre commutative maximale $\mathfrak{h}$ *(sous-algèbre de Cartan)* telle que la relation $\mathrm{ad}(X)(\mathfrak{h}) \subset \mathfrak{h}$ entraîne $X \in \mathfrak{h}$. En étudiant la représentation adjointe $H \mapsto \mathrm{ad}(H)$ de $\mathfrak{h}$ dans l'espace vectoriel $\mathfrak{g}$, on arrive alors à décomposer $\mathfrak{g}$ en somme directe de $\mathfrak{h}$ et de sous-espaces $\mathbf{C}X_\alpha$ de dimension 1, où les X_α vérifient les relations (13) à (16). On voit aisément que l'espace vectoriel réel $\mathfrak{u}$ engendré par les iH_α, les $X_\alpha - X_{-\alpha}$ et les $i(X_\alpha + X_{-\alpha})$ *est* une algèbre de Lie *réelle* dans laquelle la forme de Killing est négative non dégéné-

rée ; donc $\mathfrak{u}$ est l'algèbre de Lie d'un groupe compact semi-simple U, et $\mathfrak{g} = \mathfrak{u} \oplus i\mathfrak{u}$. Les iH_α engendrent une sous-algèbre (réelle) commutative maximale t de $\mathfrak{u}$ (correspondant à un tore maximal T de U), et on a $\mathfrak{h}$ = t $\oplus$ it.

Si l'on choisit une base (oc,), avec $1 \leqslant j \leqslant m$, du système des racines de $\mathfrak{g}$, la sous-algèbre (complexe) $\mathfrak{n}^+$ (resp. $\mathfrak{n}^-$) de $\mathfrak{g}$ ayant pour base les X_α pour $\alpha > 0$ (resp. $\alpha < 0$) est une sous-algèbre *nilpotente* : on a :

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{n}^+ \oplus \mathfrak{n}^-,$$

et $\mathfrak{b} = \mathfrak{h} \oplus \mathfrak{n}^+$ est une sous-algèbre *résoluble maximale* de g. Si G est un groupe de Lie (complexe) connexe d'algèbre de Lie $\mathfrak{g}$ et B le sous-groupe connexe de G correspondant à $\mathfrak{b}$, B est donc un sous-groupe *résoluble connexe maximal* de G. Les sous-groupes ayant ces trois propriétés sont appelés *sous-groupes de Borel* de G ; ils sont tous *conjugués* dans G. On montre que B est son propre normalisateur dans G, et que l'espace homogène G/B est *compact* et peut canoniquement être muni d'une structure de *variété algébrique projective* sur C. En outre, les doubles classes BsB forment une partition de G qui est canoniquement indexée par le *groupe de Weyl* W de U (*décomposition de Bruhat*) : de façon précise, pour tout $w \in W$, il existe, dans le normalisateur de T dans U, un élément n_w tel que $\mathrm{Ad}(n_w)$ laisse stable $\mathfrak{h}$ et induise sur $\mathfrak{h}$ la contragrédiente de w considéré comme opérant dans $\mathfrak{h}^*$ (cf. chap. 6) ; l'application $w \mapsto Bn_wB$ est une bijection de W sur l'ensemble des doubles classes modulo B.

Dans $SL(n, \mathbf{C})$, par exemple, un groupe de Bore1 est le groupe trigonal large supérieur $T(n, C)$ (cf. chap. 1).

Le groupe de Bore1 permet de donner une expression explicite de la représentation linéaire de G correspondant à une représentation linéaire de $\mathfrak{g}$ de poids dominant ω. Supposons, pour simplifier, G simplement connexe, et soit M le sous-groupe connexe de G correspondant à $\mathfrak{h}$, qui est isomorphe à $(\mathbf{C}^*)^m$ (« groupe de type multiplicatif ») ; on déduit de ω un homomorphisme ψ_ω : M $\rightarrow$ C* défini par :

$$\psi_\omega(\exp(H)) = \exp 2i\pi\omega(H),$$

coïncidant dans T avec le caractère χ_ω ; si N^+ est le sous-groupe de B correspondant à $\mathfrak{n}^+$, on a $B = M N^+$ et on prolonge ψ_ω en un homomorphisme de B dans C* en lui donnant la valeur 1 dans N^+. Soit alors V_ω l'espace vectoriel des fonctions f holomorphes dans G et vérifiant l'identité :

$$f(xb) = \psi_\omega(b)f(x), \tag{21}$$

pour $x \in G$ et $b \in B$. On peut faire opérer linéairement G dans V_ω en posant :

$$(s.f)(x) = f(s^{-1}x), \tag{22}$$

pour s, x dans G. On prouve que V_ω est de dimension finie, que la représentation de G dans Vo ainsi définie est irréductible et que sa contragrédiente a pour caractère χ_ω.

On dit qu'une algèbre semi-simple $\mathfrak{g}_0$ sur R est une *forme réelle* de $\mathfrak{g}$ si $\mathfrak{g}$ est isomorphe à la complexifiée :

$$\mathfrak{g}_0 \otimes_{\mathbf{R}} \mathbf{C}$$

de $\mathfrak{g}_0$; il est immédiat qu'il revient au même de dire que $\mathfrak{g}_0$ est isomorphe à une sous-algèbre de Lie réelle de $\mathfrak{g}$ ($\mathfrak{g}$ étant considérée comme algèbre de Lie réelle de dimension $2n$), formée des éléments invariants d'une *conjugaison* de $\mathfrak{g}$, c'est-à-dire une application semi-linéaire bijective σ de $\mathfrak{g}$ sur elle-même qui préserve le crochet, telle que $\sigma^2 = 1$. Tout revient donc à déterminer ces conjugaisons, à automorphismes de $\mathfrak{g}$ près. On montre d'abord que,

par un automorphisme de $\mathfrak{g}$, on peut ramener l'algèbre $\mathfrak{u}$ à être *stable* par σ. Alors $\mathfrak{t}_0 = \mathfrak{g}_0 \cap \mathfrak{u}$ est la sous-algèbre de $\mathfrak{u}$ formée des points fixes de la restriction de σ, et $\mathfrak{u}$ est somme directe de $\mathfrak{t}_0$ et du sous-espace $i\mathfrak{p}_0$ des $X \in \mathfrak{u}$ tels que $\sigma(X) = -X$. D'autre part, $\mathfrak{u}$ est l'ensemble des points fixes d'une conjugaison τ de $\mathfrak{g}$, qui permute à σ ; on a $\mathfrak{g}_0 = \mathfrak{t}_0 \oplus \mathfrak{p}_0$ et $\mathfrak{p}_0$ est l'ensemble des $X \in \mathfrak{g}_0$ tels que $\tau(X) = -X$. On voit que la restriction à $\mathfrak{t}_0$ (resp. $\mathfrak{p}_0$) de la forme de Killing de $\mathfrak{g}_0$ est négative (resp. positive) non dégénérée ; une telle décomposition $\mathfrak{g}_0 = \mathfrak{t}_0 \oplus \mathfrak{p}_0$ ayant toutes ces propriétés est appelée *décomposition de Cartan* de la forme réelle $\mathfrak{g}_0$ de $\mathfrak{g}$; elle est déterminée à un automorphisme près de la forme Ad(s).

Une fois connu un groupe semi-simple compact U, la détermination des formes réelles de son complexifié G revient donc à la détermination des *automorphismes involutifs* s de U (dont σ, restreint à $\mathfrak{u}$, sera l'automorphisme dérivé s_*). Si K est le sous-groupe compact de U formé des x invariants par s, sa composante connexe K_0 correspond à $\mathfrak{t}_0$; d'autre part, si P est la composante connexe de e dans l'ensemble des $x \in U$ tels que $s(x) = x^{-1}$ (qu'on montre être égale à $\exp(i\mathfrak{p}_0)$), on a $U = K_0 P$; le groupe U agit sur P transitivement par :

$$(u, p) \mapsto ups(u^{-1}),$$

et P est isomorphe à l'espace riemannien symétrique U/K. En outre, l'application $(k, p) \mapsto kp$ restreinte à un ouvert partout dense convenable de $K_0 \times P$, est un difféomorphisme sur un ouvert partout dense de U. Si A désigne un sous-groupe commutatif maximal de P (A est ici un tore), on prouve que tout $p \in P$ s'écrit kak^{-1} pour un $a \in A$ et $k \in K_0$, d'où $U = K_0AK_0$ (*décomposition de Cartan*).

Par exemple, si U = SO(3, R), groupe des rotations de $\mathbf{R}^3$, on peut prendre pour s l'automorphisme $t \mapsto rtr^{-1}$, où r est une réflexion orthogonale ; alors K = O(2, R) et K_0 = SO(2, R); A est un tore de dimension 1, comme K_0, et la décomposition K_0AK_0 n'est autre que la classique description des rotations de $\mathbf{R}^3$ par les trois *angles d'Euler*.

De la même manière, si G_0 est la forme réelle (non compacte) correspondant à σ, c'est-à-dire le sous-groupe connexe de G correspondant à l'algèbre de Lie $\mathfrak{g}_0$, la restriction de τ à $\mathfrak{g}_0$ est de la forme t_*, où t est un automorphisme involutif de G_0 ; K_0 est l'ensemble des $x \in G_0$ invariants par t, et $P_0 = \exp(\mathfrak{p}_0)$ la composante connexe de e dans l'ensemble des $x \in G_0$ tels que $t(x) = x^{-1}$; on obtient $G_0 = K_0 P_0$, et ici l'application $(k, p) \mapsto kp$ de $K_0 \times P_0$ sur G_0 est un difféomorphisme. L'espace P est isomorphe à un $\mathbf{R}^n$; G_0 y agit transitivement par $(x, p) \mapsto xpt(x^{-1})$, et, pour cette action, P_0 est isomorphe à l'espace riemannien symétrique G_0/K_0. On a encore la décomposition de Cartan $G_0 = K_0A_0K_0$ (A, sous-groupe commutatif connexe maximal de P_0), et, en analysant de plus près la représentation adjointe de A dans $\mathfrak{g}_0$, on obtient la décomposition d'Iwasawa $G_0 = K_0A_0N$ décrite dans le chapitre 2.

On voit notamment que les espaces riemanniens symétriques compacts simplement connexes sont en *correspondance biunivoque* avec les espaces riemanniens symétriques non compacts.

8. Représentations linéaires de dimension infinie

La description des représentations irréductibles d'un groupe semi-simple complexe donnée dans (21) et (22) est un exemple

particulier de l'idée fondamentale de *représentation linéaire induite,* initialement introduite par Frobenius pour les groupes finis, (cf. la partie D ci-dessus Représentation linéaire des groupes), appliquée aux groupes de Lie.

D'une façon générale, soit G un groupe de Lie, Γ un sous-groupe fermé de G, F un espace vectoriel complexe de dimension finie, et $\xi \mapsto L(\xi)$ une représentation linéaire de Γ dans F. Soit alors V un espace vectoriel (en général de dimension *infinie*) de *fonctions* définies dans G et vérifiant l'identité :

$$f(x\xi) = L(\xi) \cdot f(x) \tag{23}$$

pour $x \in$ G et $\xi \in \Gamma$. On fait alors opérer G dans V en posant :

$$(s \cdot f)(x) = f(s^{-1}s) \tag{24}$$

pour s, x dans G.

L'exemple des représentations de dimension finie donné dans (21) et (22) correspond au cas où L est une représentation de dimension 1 et où V est de dimension finie. Le cas le plus étudié en dehors de ce dernier est celui où L est une représentation unitaire (autrement dit, L(E,) laisse invariant un produit scalaire euclidien dans F) ; si f'vérifie (23), on a $\| f(x\xi) \| = \|f(x)\|$ pour la norme euclidienne dans F, et on peut considérer f comme définie dans G/Γ ; on définit alors V comme l'espace de Hilbert des fonctions f sur G/Γ telles que :

$$\int_{G/\Gamma} \|f(z)\|^2 d\mu(z) < +\infty, \tag{25}$$

où μ est une mesure sur GΓ invariante pour l'action de G. Il s'agit de savoir si cette représentation est irréductible, ou de la décomposer en représentations irréductibles ; cela pose des problèmes difficiles qui sont encore loin d'être tous résolus. Leur intérêt réside dans le fait qu'ils rattachent à la théorie des groupes de Lie des questions d'analyse ou de physique d'allure toute différente.

En premier lieu, on rencontre ainsi de façon naturelle de nombreuses *fonctions spéciales,* dont on peut ainsi faire une théorie unifiée et « expliquer » maintes propriétés qui paraissaient fortuites.

Les exemples les plus simples s'obtiennent lorsqu'on prend G = U. groupe compact semi-simple, et $\Gamma = K_0$ (notations du chap. 7). Alors toutes les représentations de U dans un espace de Hilbert V se décomposent en représentations de dimension finie, le groupe U opérant pour ces représentations dans des sous-espaces de dimension finie V_j de V, dont la somme est directe et partout dense. *É.* Cartan s'est le premier aperçu que les fonctions constituant les V, ont des propriétés remarquables. Pour U = SO(3, R), K = SO(2, **R**), par exemple, on obtient ainsi *les fonctions sphériques* classiques définies sur $U/K = \mathbf{S}_2$ (sphère à 2 dimensions) comme restrictions des polynômes harmoniques homogènes dans $\mathbf{R}^3$.

On obtient d'autres fonctions spéciales, telles que les fonctions de Bessel ou les fonctions hypergéométriques, en prenant pour G certains groupes de dimension 4.

Si on prend pour G un groupe semi-simple et pour Γ un sous-groupe *discret* convenable, on obtient cette fois comme fonctions « spéciales » ce qu'on appelle des fonctions (ou formes) *automorphes*, qui constituent une vaste généralisation des « fonctions fuchsiennes » de H. Poincaré.

Les fonctions appartenant à V ne sont pas nécessairement continues dans G, mais on peut montrer qu'il y a toujours un sous-espace dense V_0 de V, stable pour

la représentation de G et tel que, pour tout X∈ $\mathfrak{g}$ (algèbre de Lie de G), la dérivée pour $t = 0$ de l'application $t \mapsto \exp(tX) f$ existe pour tout $f \in V_0$. Si on note ρ(X) . f cette limite, ρ(X) devient un opérateur linéaire de l'algèbre de Lie $\mathfrak{g}$ dans l'espace (de dimension infinie en général) V_0.

En physique quantique, V est un espace de « fonctions d'ondes », et G est soit le « groupe de Poincaré » (produit semi-direct du groupe de Lorentz SO(3, 1) et du groupe commutatif $\mathbf{R}^4$), soit (dans la théorie récente des « particules élémentaires ») le produit de ce groupe et d'un groupe compact tel que SU(2), SU(3) ou SU(6). Les opérateurs ρ(X) sont des opérateurs différentiels du premier ordre ; on peut étendre la représentation $X \mapsto \rho(X)$ de $\mathfrak{g}$ à une représentation de l'algèbre enveloppante $\mathfrak{G}$ (cf. chap. 5) ; pour k éléments $X_j \in \mathfrak{g}$,

$$\rho(X_1 X_2 \dots X_k)$$

sera le produit d'opérateurs :

$$\rho(X_1)\rho(X_2) \dots \rho(X_k) ;$$

ce sont donc des opérateurs différentiels d'ordre quelconque. Un intérêt particulier s'attache aux opérateurs ρ(Z) (dits *opérateurs de Casimir*) où Z appartient au *centre* de $\mathfrak{G}$; comme ils *commutent* avec les ρ(X), ils agissent par *homothétie* dans chaque sous-espace V_j de V où la restriction de la représentation considérée de G est *irréductible* ; en d'autres termes, les fonctions $f \in V_j$ satisfont à des équations aux dérivées partielles $\rho(Z)f = \lambda f$. Ainsi, pour G = SO(3, R), le centre de $\mathfrak{G}$ est engendré par l'unique élément $Z = X_1^2 + X_2^2 + X_3^2$, où (X_j), $j = 1, 2, 3$ est une base convenable de l'algèbre de Lie $\mathfrak{so}(3, \mathbf{R})$, et on constate que **ρ(Z)** est le *laplacien*, ce qui « explique » que l'on obtienne comme éléments des V, des polynômes harmoniques. En physique quantique, les *valeurs propres A* des opérateurs **ρ(Z)** dans chaque V_j sont mises en correspondance avec les valeurs des grandeurs physiques fondamentales (tels que masse, spin, isospin, etc.) de la « particule élémentaire » associée à la représentation irréductible dans V_j

9. Généralisations

On constate que les groupes semi-simples complexes sont des *groupes linéaires algébriques*, c'est-à-dire des sous-groupes G de groupes linéaires $\mathbf{GL}(n, \mathbf{C})$, définis par des *équations algébriques* entre les éléments des matrices de G. On sait d'autre part que les groupes classiques peuvent être aussi définis pour un corps de base K quelconque au lieu du corps C (cf. la partie B ci-dessus - Groupes classiques et géométrie). On est donc conduit à se demander s'il n'existe pas une « théorie de Lie » pour les groupes linéaires sur un corps quelconque K et, comme ici il n'y a plus de notions topologiques, on les remplace par la restriction que les groupes considérés sont *algébriques* au sens ci-dessus.

On peut alors développer toute une théorie dont les résultats (mais non les méthodes) se calquent sur ceux de la théorie des groupes de Lie (Borel-Chevalley). On définit une notion (algébrique) de « connexion » et des notions telles que celle de radical, de groupe semi-simple ou de sous-groupe de Borel d'un groupe algébrique exactement comme pour les groupes de Lie. Le résultat le plus remarquable est que, lorsque le corps de base K est algébriquement clos (mais de caractéristique quelconque), les groupes semi-simples sont encore donnés par la classi-

fication de Killing-Cartan. Il n'y a plus ici de méthode « infinitésimale » à proprement parler, bien qu'on puisse encore définir une algèbre de Lie $\mathfrak{g}$ (et même une algèbre associative $\mathfrak{G}$) associée à un groupe linéaire algébrique G ; mais son utilité est bien moindre que dans la théorie classique. Les raisonnements essentiels sont de nature globale et reposent sur le fait que, pour un sous-groupe de Bore1 B de G, le quotient G/B est encore muni d'une structure de *variété algébrique projective*. En outre, on étend encore à ce cas la décomposition de Bruhat (cf. chap. 7), qui joue également un rôle important dans les démonstrations.

Aux groupes linéaires algébriques définis sur le corps des rationnels Q est maintenant rattachée la théorie des *groupes arithmétiques* : si G est un sous-groupe algébrique de $\mathbf{GL}(n, \mathbf{Q})$, on dit qu'un sous-groupe Γ de G est arithmétique s'il laisse stable un *réseau*, c'est-à-dire un sous-Z-module de $\mathbf{Q}^n$ engendré par une base de $\mathbf{Q}^n$. Par exemple, $\mathbf{SL}(n, Z)$ est un groupe arithmétique dans $\mathbf{SL}(n, \mathrm{Q})$. En considérant G et Γ comme sous-groupes du groupe de Lie $G_{\mathbf{R}}$ (ensemble des matrices de $\mathbf{GL}(n, \mathrm{R})$ vérifiant les mêmes équations algébriques que celles qui définissent G) et en utilisant à fond les techniques de la théorie des groupes semi-simples (notamment les décompositions de Bruhat et d'Iwasawa), on retrouve la théorie de la « réduction » des formes quadratiques à coefficients entiers de Hermite-Minkowski et le théorème de finitude de Jordan sur les classes de formes de degré $\geqslant 3$ à coefficients entiers, et on les généralise considérablement.

Au lieu de considérer G comme plongé dans $G_{\mathbf{R}}$, on peut aussi le considérer comme plongé dans $G_{\mathbf{Q}_p}$, où $\mathbf{Q}_p$ est le corps des *nombres p-adiques*. Comme $\mathbf{Q}_p$ est ici muni d'une topologie, la correspondance entre algèbre de Lie et groupe de Lie est presque aussi satisfaisante en théorie p-adique que pour les groupes de Lie réels ou complexes. En utilisant à la fois les plongements de G dans $G_{\mathbf{R}}$ et dans les $G_{\mathbf{Q}_p}$ correspondant à *tous* les nombres premiers p, on arrive aux résultats les plus profonds de Minkowski-Siegel sur les formes quadratiques à coefficients entiers, ici encore largement généralisés et placés dans leur cadre naturel. Il est intéressant de noter que, dans cette théorie, un rôle particulièrement important est tenu par les *mesures invariantes* sur les groupes p-adiques.

Enfin, la théorie des groupes semi-simples est liée de façon inattendue à celle des groupes finis. Si l'on part d'une base de Weyl-Chevalley (cf. chap. 6) d'une algèbre de Lie semi-simple complexe $\mathfrak{g}$, on constate que, pour chaque α, l'application :

$$t \mapsto \exp(t\, \mathrm{ad}(\mathrm{X}_\alpha))$$

est un isomorphisme du groupe additif C sur un groupe $\mathrm{X}\alpha$ de matrices qui, en vertu des relations (16), se trouvent avoir des éléments qui sont des *polynômes en t à coefficients entiers*. Cela permet de définir ces matrices lorsque t appartient à un corps *(commutatif) arbitraire* K. En prenant pour α toutes les racines de $\mathfrak{g}$, on obtient dans $\mathbf{GL}(n, \mathrm{K})$ (où $n = \dim \mathfrak{g}$) un ensemble de matrices qui engendrent un sous-groupe G_K de $\mathbf{GL}(n, \mathrm{K})$, appelé *groupe de Chevalley* sur K associé à $\mathfrak{g}$. En particulier si K est un corps fini, le groupe G_K est un groupe *fini*.

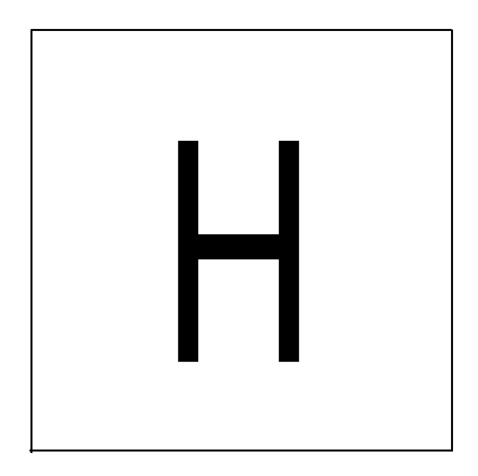
On prouve alors que, si l'on part d'une algèbre de Lie *simple* complexe $\mathfrak{g}$, le groupe de Chevalley correspondant G_K est *simple* (au sens de la théorie des groupes « abstraits ») sauf dans quatre cas correspondant à des corps à deux ou trois éléments

et à des algèbres de l'un des types A,, B_2 ou G_2. Ces groupes simples, ont, en outre, des décompositions de Bruhat qui permettent d'étudier de façon détaillée leur structure et qui (pour le cas d'un corps K fini) les distinguent nettement des autres types de groupes simples finis.

JEAN DIEUDONNÉ

Bibliographie

H. BACRY, *Leçons sur la théorie des groupes et les symétries des particules élémentaires*, Dunod, Paris, 1968 / A. BOREL, *Linear Algebraic Groups*. repr. of 1969, Springer-Verlag, New York, 2e éd. 1991 ; *Introduction aux groupes arithmétiques*, Hermann, Paris, 1969 / N. BOURBAKI, *Groupes et algèbres de Lie*, 5 vol., Masson, 1982 / J. DIEUDONNÉ. *Special Functions and Linear Representations of Lie Groups*. American Mathematical Society, Providence (R.I.), 1 982 / I. CELFAND, *Representation Theory Selected Papers*, Cambridge Univ. Press, 1982 / G. HOCHSCHILD, *Basic Theory of Algebraic Groups and Lie Algebras*, Springer-Verlag, New York, 1981 / N. JACOBSON, *Lie Algebras*, repr. of 1962, Dover Publ., 1979 / R. MNEIMÉ & F. TESTARD, *Introduction à la théorie des groupes de Lie classiques*, Hermann, 1986 / V. S. VARADARAJAN, *Lie Groups, Lie Algebras and their Representations*, Springer-Verlag, *1914*, rééd. 1988 / N. VILENKIN, *Special Functions and the Theory of Group Representations*, repr. of 1983 ; *American Mathematical Society*, Providence (R.I.), 1988 / B. G. WYBOURNE, *Classical Groups for Physicists*, repr. of 1974, Books on Demand, Ann Arbor (Mich.) / 0. L. WEAVER & D. H. SATTINGER, *Lie Groups and Algebras with Applications to Physics, Geometry and Mechanics*, Springer, 1986.



HARMONIQUE ANALYSE

Lorsqu'on fait vibrer, dans des conditions idéales, une corde de longueur l, fixée en ses extrémités d'abscisses 0 et l, l'équation aux dérivées partielles :

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

est vérifiée, où $u(x, t)$ est une fonction dont la valeur représente, à l'instant t, le déplacement transversal, par rapport à la position d'équilibre, du point d'abscisse x.

D'Alembert donne, en 1747, la solution de cette équation sous la forme :

$$u(x, t) = f(ct + x) - f(ct - x),$$

où f est une fonction quelconque de période 2 l. Quelques années plus tard, en 1753, Daniel Bernoulli considère des solutions particulières de l'équation des cordes vibrantes, de la forme :

$$2 \sin \frac{n\pi}{l} x \cos \frac{n\pi}{l} ct,$$

ou encore :

$$\sin\left[\frac{n\pi}{l}(ct + x)\right] - \sin\left[\frac{n\pi}{l}(ct - x)\right],$$

pour toute valeur entière positive de n. Ces solutions correspondent aux fonctionsfde la forme :

$$f(x) = \sin \frac{n\pi}{l} x.$$

Or les fonctions trigonométriques :

$$\sin \frac{n\pi}{l} x \quad \text{ou} \quad \cos \frac{n\pi}{l} x$$

sont les plus simples des fonctions de période 2 l. D'où l'idée, avancée par Bernoulli, que la fonction f la plus générale, qui intervient dans la solution de d'Alembert, peut s'exprimer sous la forme d'une série trigonométrique :

$$f(x) = \frac{a_0}{2} + \sum_{n=1} (a_n \cos \frac{n\pi}{l} x + b_n \sin \frac{n\pi}{l} x),$$

ou, de manière équivalente :

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a, \sin \frac{n\pi}{l} (x - \beta_n).$$

Le terme correspondant à $n = 1$ donne alors la vibration fondamentale de la corde, les termes suivants correspondent aux harmoniques (cela rejoint l'expérience acoustique courante) ; de plus, le coefficient α_n régit l'intensité de l'harmonique d'ordre n, et β_n en définit la phase.

Ainsi le problème des cordes vibrantes menait tout naturellement à la question suivante : une fonction périodique peut-elle se représenter par une série trigonométrique ? L'analyse harmonique classique est, cn principc, la branche des mathématiqucs qui traite de problèmes de ce type.

Pour obtenir des éléments de réponse à cette question fondamentale, il a fallu, à partir du milieu du XVIIIe siècle, que les mathématiciens se fassent une idée de plus en plus précise des objets sur lesquels ils travaillaient. C'est ainsi que l'étude de la représentation des fonctions périodiques par des séries trigonométriques devait fortement contribuer à la prise de conscience de la notion de fonction : la conception moderne d'une fonction, définie comme une correspondance, et pouvant fort bien ne posséder aucune des propriétés usuelles de régularité (continuité, dérivabilité, intégrabilité), émergea peu à peu lorsqu'il devint évident que l'idée naïve d'une fonction donnée par une formule explicite était insuffisante : il fallut tout à la fois préciser ce qu'on entendait par « fonction quelconque » et considérer des classes particulières de fonctions dont les propriétés spéciales, soigneusement mises en évidence, permettaient de résoudre un problème donné.

Ensuite, la théorie des distributions et celle des groupes topologiques sont venues proposer diverses directions dans lesquelles l'analyse harmonique se généralise et s'approfondit ; celle-ci est devenue une branche importante des mathématiques, en relation avec les distributions, les algèbres normées, les probabilités, les espaces de Hilbert, les fonctions de variable complexe et s'est étendue aux fonctions non linéaires.



1. Les séries de Fourier

Les coefficients de Fourier

Considérons une fonction f à valeurs réelles ou complexes, d'une variable réelle, périodique, de période 2π pour fixer les idées. Si f admet un développement **en** série trigonométrique :

$$(1) \quad f(x) = ? + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

et que la série $\sum(|a_k| + |b_k|)$ soit convergente, on peut intégrer terme à terme, entre 0 et 2π, les séries :

$$f(x)\cos nx = \frac{a_0}{2}\cos nx + \sum_{k=1}^{\infty}(a_k \cos kx \cos nx + b_k \sin kx \cos nx),$$

$$f(x)\sin nx = \frac{a_0}{2}\sin nx + \sum_{k=1}^{\infty}(a_k \cos kx \sin nx + b_k \sin kx \sin nx).$$

Compte tenu des relations, valables pour des entiers n et k :

$$\int_0^{2\pi}\cos nx \cos kx\,dx = \begin{cases} 0 & \text{si } k \neq n, \\ \pi & \text{si } k = n, k \neq 0, \\ 2\pi & \text{si } k = n = 0 \end{cases}$$

$$\int_0^{2\pi}\sin nx \cos kx\,dx = 0, \ \forall\, k \text{ et } n,$$

$$\int_0^{2\pi}\sin nx \sin kx\,dx = \begin{cases} 0 & \text{si } k \neq n, \\ \pi & \text{si } k = n, \end{cases}$$

on obtient les valeurs des coefficients a_n et b_n directement à partir de la somme $f(x)$ de la série donnée :

$$(2) \qquad \begin{cases} a_n = \dfrac{1}{\pi}\displaystyle\int_0^{2\pi} f(x)\cos nx\,dx, \\ b_n = \dfrac{1}{\pi}\displaystyle\int_0^{2\pi} f(x)\sin nx\,dx\,; \end{cases}$$

ce sont les formules de Fourier.

Si, maintenant, on part d'une fonction f, de période 2π, continue (il suffit, en fait, qu'elle soit intégrable, au sens de Lebesgue sur $[0, 2\pi]$), il est naturel de considérer, par analogie avec ce qui précède, les coefficients a_n et b_n, donnés, pour un entier $n \geqslant 0$, par les formules (2). Ce sont les *coefficients de Fourier* de la fonction f, et la série qu'ils définissent est la *série de Fourier* de f. Rien ne permet de préjuger de la convergence de cette série vers f, aussi la relation entre f et sa série de Fourier n'est-elle pas notée par le signe d'égalité, mais on écrit :

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx).$$

Si, au lieu d'une fonction de période 2π, on considère une fonction f de période T, on définit de manière analogue les coefficients de Fourier de f par les formules :

$$a_n = \frac{2}{T}\int_0^T f(x)\cos\frac{2\pi}{T}nx\,dx,$$

$$b_n = \frac{2}{T}\int_0^T f(x)\sin\frac{2\pi}{T}nx\,dx\,;$$

la série de Fourier de f est alors :

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{2n\pi}{T}x + b_n \sin\frac{2n\pi}{T}x.$$

Questions de convergence

Le problème de la représentation d'une fonction périodique par une série trigonométrique se ramène à l'étude de la convergence de sa série de Fourier. Nous nous contenterons de donner ici quelques-uns des nombreux résultats obtenus dans ce domaine (cf. SÉRIES TRIGONOMÉTRIQUES).

a) D'abord, en dehors de toute notion de convergence, la série de Fourier d'une fonction caractérise celle-ci (cela doit être compris comme une caractérisation en tant que fonction mesurable au sens de Lebesgue, deux fonctions étant considérées comme équivalentes lorsque l'ensemble des points où elles diffèrent est négligeable pour la mesure de Lebesgue : on dit alors qu'elles sont égales presque partout (cf. INTÉGRATION ET MESURE) ; si l'on se restreint à la classe des fonctions continues. l'égalité presque partout entraîne l'égalité partout). Autrement dit, si deux fonctions ont les mêmes séries de Fourier, elles sont égales presque partout.

b) Si les coefficients de Fourier, a_n et b_n, d'une fonction continue f forment une série absolument convergente (c'est-à-dire si la série $\Sigma(|a_n| + |b_n|)$ converge), alors la série de Fourier de f converge uniformément vers une fonction continue qui, d'après ce qui précède, a même série de Fourier que f. Le résultat ci-dessus montre donc quefest égale à la somme de sa série de Fourier.

Certains critères permettent d'affirmer qu'une fonction continue f possède la propriété ci-dessus. Montrons que c'est le cas si, par exemple, la fonctionfadmet une dérivée seconde continue. Soit a, et b_n les coefficients de Fourier de f, a'_n et b'_n ceux de f', a'', et h'',, ceux def'' (f' etf'' sont respectivement les dérivées première et seconde def). Une intégration par parties dans les intégrales (2) écrites pour f' montre que l'on a :

$$a'_n = nb_n,\ b'_n = -na_n,$$

et, par suite :

$$a''_n = -n^2 a_n,\ b''_n = -n^2 b_n.$$

Or (2) entraîne que les coefficients de Fourier d'une fonction sont bornés en module (ils sont majorés par l'intégrale sur $[0, 2\pi]$ du module de cette fonction, divisée par π), de sorte que la suite $n^2(|a_n| + |b_n|)$ est bornée, ce qui montre que la série $\Sigma(|a_n| + |b_n|)$ converge. On obtient ainsi le résultat annoncé, qu'une fonction deux fois continûment dérivable est la somme de sa série de Fourier, la convergence étant d'ailleurs absolue et uniforme.

c) Un résultat plus profond, dû à Dirichlet et à Jordan, est le suivant, que nous donnons dans un cas particulier : Si la fonction f, continue et périodique de période 2π, possède une dérivée continue, sa série de Fourier converge uniformément vers f.

d) Si l'on fait simplement l'hypothèse quefest continue, on ne peut plus affirmer que la série de Fourier de f converge vers f, ce qui signifierait que la suite des fonctions :

$$S_m(x) = \frac{a_0}{2} + \sum_{n=1}^{m} (a_n \cos nx + b_n \sin nx)$$

tend vers f lorsque m augmente indéfiniment. Au lieu des sommes partielles S_m de la série de Fourier de f, considérons les fonctions :

$$\sigma_m(x) = \frac{S_1(x) + S_2(x) + \dots + S_m(x)}{m}$$

appelées *moyennes de Césaro* de la fonction f. On a alors le théorème de Fejer : Si f est continue, ses moyennes de Césaro tendent uniformément vers f.

Coefficients de Fourier exponentiels

Si l'on introduit la fonction exponentieile complexe e^{ix}, liée aux fonctions trigonométriques par les relations d'Euler :

$$e^{ix} = \cos x + i \sin x,$$
$$\cos x = \frac{e^{ix} + e^{-ix}}{2}$$
$$\sin x = \frac{e^{ix} - e^{-ix}}{2i}$$

on obtient l'égalité :

$$a_n \cos nx + b_n \sin nx = c_n e^{inx} + c_{-n} e^{-inx},$$

où les coefficients a_n, b_n, c_n, c_{-n} vérifient les relations :

$$c_n = \frac{a_n - ib_n}{2}$$
$$c_{-n} = \frac{a_n + ib_n}{2}$$
$$a_n = c_n + c_{-n},$$
$$b_n = i(c_n - c_{-n}).$$

Si f est périodique et intégrable, on appellera coefficients de Fourier exponen-

tiels les nombres c_n, définis pour tout entier *relatif* n, par :

$$(3) \qquad c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx}\, dx,$$

et on appellera encore *série de Fourier* de f (cf. SÉRIES TRIGONOMÉTRIQUES) la série :

$$\sum_{n=-\infty}^{+\infty} c_n e^{inx}.$$

Moyennant les relations ci-dessus entre les a_n, b_n ($n \geqslant 0$) et c_n (n quelconque), il y a identité formelle entre cette notion de série de Fourier et la notion antérieure de série de Fourier trigonométrique (où l'on peut introduire un coefficient b_0 égal à 0).

Nous utiliserons désormais la forme « exponentielle » de la série de Fourier d'une fonction f :

$$f(x) \sim \sum_{n=-\infty}^{+\infty} c_n e^{inx},$$

qui donne des calculs et des formules plus simples, et correspond mieux à la nature profonde de la situation mathématique, comme nous le verrons au chapitre 4.

Le théorème de Bessel-Parseval-Plancherel

Soit f une fonction périodique continue, et soit :

$$\sum_{n=-\infty}^{+\infty} c_n e^{inx},$$

sa série de Fourier. On a alors l'égalité suivante :

$$(4) \qquad \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2\, dx = \sum_{n=-\infty}^{+\infty} |c_n|^2.$$

Plus généralement, si on considère une fonction périodique, de carré intégrable sur $[0, 2\pi]$, cette fonction est en particulier intégrable, et possède des coefficients de Fourier c_n tels que l'égalité (4) ait lieu.

Tout aussi remarquable est le fait que toute suite (c_n), n parcourant l'ensemble des entiers relatifs, telle que :

$$\sum_{n=-\infty}^{+\infty} |c_n|^2$$

converge, est la suite des coefficients de Fourier d'une fonction de carré intégrable.

En d'autres termes, la correspondance entre une fonction de carré intégrable sur $[0, 2\pi]$ et la suite de ses coefficients de Fourier définit un isomorphisme isométrique entre l'espace de Hilbert $L^2([0, 2\pi])$ et l'espace de Hilbert l^2 des suites de carré intégrable (les structures hilbertiennes des deux espaces ci-dessus sont définies par les produits scalaires :

$$(f \mid g) = \frac{1}{2\pi} \int_0^{2\pi} f(x) \overline{g(x)}\, dx,$$

$$(\{c_n\} \mid \{d_n\}) = \sum_{n=-\infty}^{+\infty} c_n \overline{d_n},$$

la convergence de l'intégrale et de la série étant assurée par l'inégalité de Schwarz ; cf. INTÉGRATION ET MESURE, espace de HILBERT).

Pour les fonctions périodiques de carré intégrable sur $[0, 2\pi]$, les méthodes hilbertiennes simplifient l'étude des séries de Fourier. Citons par exemple le résultat suivant, dont la démonstration est d'ailleurs facile : Si f est périodique, de carré intégrable sur $[0, 2\pi]$, la série de Fourier :

$$\sum_{n=-\infty}^{+\infty} c_n e^{inx}$$

defconverge vers$f$au sens de l'espace L^2 ; cela signifie que :

$$\lim_{\substack{p \to -\infty \\ q \to +\infty}} \left[\int_0^{2\pi} \left| f(x) - \sum_p^q c_n e^{inx} \right|^2 dx \right] = 0.$$

Les fonctions presque-périodiques

La somme de deux fonctions périodiques dont les périodes sont dans un rapport irrationnel n'est pas périodique (par exemple, $e^{ix} + e^{-i\pi x}$, sin x+cos $\sqrt{2}x$ sont des modèles de telles fonctions). Cependant, les fonctions de ce type ont des propriétés voisines de la périodicité. Cette idée a conduit Harald Bohr, vers 1925, à la notion de fonction presque-périodique.

Une fonction f, continue sur la droite réelle, est appelée *presque-périodique* si, pour tout $\varepsilon > 0$, il existe un nombre T > 0 tel que tout intervalle de longueur T contienne au moins un nombre c tel que l'on ait :

$$|f(x+c) - f(x)| < \varepsilon$$

pour toute valeur réelle de la variable x. Une fonction périodique possède évidemment cette propriété. De nombreuses définitions équivalentes de la presque-périodicité peuvent être données ; en voici une : Une fonction est presque-périodique si, et seulement si, elle peut être approchée uniformément par une suite de fonctions de la forme :

$$\sum_{n=1}^{N} c_n e^{i\lambda_n x},$$

où les c,, sont des nombres complexes et les λ_n des nombres réels.

Les fonctions presque-périodiques forment un anneau, et toute limite uniforme d'une suite de fonctions presque-périodiques est une fonction presque-périodique.

Pour toute fonction presque-périodique f, l'expression :

$$\frac{1}{2T} \int_{-T}^{T} f(x)\, dx$$

possède une limite lorsque T tend vers l'infini. Cette limite, notée $M(f)$, est appelée valeur moyenne de f.

Par le biais des moyennes, on peut définir la série de Fourier d'une fonction presque-périodique. Pour tout λ réel, la fonction $f(x)e^{-i\lambda x}$ est presque-périodique ; appelons c_λ sa moyenne. On montre que les valeurs de A telles que c_λ ne soit pas nul forment un ensemble fini ou dénombrable A,, A,, ... A,, ; on appelle série de Fourier de f l'expression :

$$\sum c_{\lambda_n} e^{i\lambda_n x} ;$$

lorsquefest périodique **de** période 2π, les A, sont entiers et on retrouve la notion usuelle de série de Fourier.

Les séries de Fourier des fonctions presque-périodiques jouissent de propriétés analogues à celles des séries de Fourier des fonctions périodiques. Citons par exemple le résultat suivant, analogue à la relation (4) : Pour toute fonction presque-périodique$f$, $|f|^2$ est presque-périodique et, si $\Sigma c_n e^{i\lambda_n x}$ est la série de Fourier de f, on a :

$$M(|f|^2) = \sum |c_n|^2.$$

2. Analyse et synthèse harmoniques

Considérons une fonction f continue, de période 2π, et soit :

$$\sum_{n=-\infty}^{+\infty} c_n e^{inx}$$

sa série de Fourier. L'égalité :

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int} \, dt,$$

entraîne :

$$c_n e^{inx} = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{in(x-t)} \, dt$$
$$= \frac{1}{2\pi} \int_0^{2\pi} f(x-t) e^{int} \, dt$$

Si nous appelons f_t la translatée de f par t, définie par $f_t(x) = f(x - t)$, nous obtenons :

$$c_n e^{inx} = \frac{1}{2\pi} \int_0^{2\pi} f_t(x) e^{int} \, dt.$$

En considérant l'intégrale comme une limite de sommes finies, on peut dire, de manière peu précise mais imagée, que $c_n e^{inx}$ (ou aussi bien e^{inx} lorsque c_n n'est pas nul) est limite de combinaisons linéaires de translatées de f.

Cela nous conduit à la notion d'analyse harmonique et de spectre d'une fonction. Soit E un espace vectoriel topologique (cf. espaces vectoriels TOPOLOGIQUES) de fonctions définies sur l'ensemble R des nombres réels, tel que si $f \in$ E et $t \in$ R, la translatée f_t appartienne à E, avec certaines conditions de continuité ; on suppose, de plus, que toute fonction exponentielle $e^{i\lambda x}$, λ réel, appartient à E.

On dira qu'un nombre réel λ appartient au *spectre* d'un élément f de E si la fonction $e^{i\lambda x}$ peut être approchée, au sens de la topologie de E, par des combinaisons linéaires de translatées de f. On note o,(f), spectre def'dans E, l'ensemble des tels λ (pour une fonction donnée, la notion de spectre peut dépendre de la topologie dont est muni l'espace E).

Le problème de l'analyse harmonique dans E d'une fonction f est la détermination de $\sigma_E(f)$. Le problème de la synthèse harmonique dans E de f est le suivant : f est-elle limite, dans E, de combinaisons linéaires d'exponentielles, $e^{i\lambda x}$, avec $\lambda \in \sigma_E(f)$? Si la réponse est positive, on dit quefsatisfait à la synthèse harmonique (ou encore que f est synthétisable) dans E.

Prenons, par exemple, pour E l'espace des fonctions continues et bornées sur R, avec la topologie de la convergence uniforme. Si f est une fonction périodique appartenant à E, le spectre de f est l'ensemble des entiers n tels que les coefficients de Fourier c_n correspondants soient non nuls ; le théorème de Fejer sur la convergence uniforme vers f des moyennes de Césaro montre que f est synthétisable dans E. Plus généralement, si f est presque-périodique, de série de Fourier :

$$\sum c_n e^{i\lambda_n x},$$

avec c,, $\neq$ 0, le spectre defest l'ensemble des exposants A, qui figurent dans la série de Fourier de f. Pour une fonction continue et bornée quelconque, le spectre est plus difficile à déterminer en général. Il est remarquable que, dans l'espace E que nous considérons ici, les fonctions synthétisables soient exactement les fonctions presque-périodiques.

En fait, les questions les plus intéressantes concernant la synthèse harmonique se posent lorsque l'on prend pour E l'espace L"(R), muni de sa topologie faible d'espace dual de $L^1(\mathbf{R})$ (cf. INTÉGRATION ET MESURE, et espaces vectoriels TOPOLOGIQUES). Ces problèmes, ainsi que leur généralisation au cas où l'on considère d'autres groupes que R (cf. chap. 4) sont loin d'être tous résolus et leur étude est l'un des points essentiels de l'analyse harmonique moderne.

3. La transformation de Fourier

Certaines classes importantes de fonctions ne se prêtent pas à l'analyse harmonique

telle qu'elle a été définie ci-dessus. Ainsi, l'espace L'(R) des (classes de) fonctions intégrables sur R ne contient aucune exponentielle ; aussi utilise-t-on un autre procédé pour en faire l'analyse et la synthèse. C'est la transformation de Fourier qui permet de définir le spectre d'une fonction intégrable et, dans certains cas, d'en faire la synthèse.

Soit f une fonction intégrable (par exemple, continue et nulle hors d'un ensemble borné). À f on associe une autre fonction définie sur R, notée fou $\mathcal{F}f$, et appelée *transformée de Fourier* de f :

$$(5) \quad \hat{f}(t) = (\mathcal{F}f)(t) = \int_{\mathbf{R}} f(x) e^{-2i\pi tx} \, dx.$$

La présence du coefficient 2π est conventionnelle (la convention n'est d'ailleurs pas universelle) et permet d'avoir une formule de réciprocité particulièrement simple.

Définissons, outre l'opérateur $\mathcal{F}$ de transformation de Fourier, l'opérateur $\bar{\mathcal{F}}$ de transformation de Fourier conjuguée (ou réciproque) :

$$(6) \qquad (\bar{\mathcal{F}}f)(t) = \int_{\mathbf{R}} f(x) e^{2i\pi tx} \, dx.$$

Pour toute fonction f, et tout réel t, on a :

$$(\bar{\mathcal{F}}f)(t) = (\mathcal{F}f)(-t).$$

Propriétés de la transformation de Fourier

a) Pour toute fonction intégrable f, $\mathcal{F}f$ est continue et tend vers 0 à l'infini. Si on désigne par A(R) l'ensemble des fonctions $\mathcal{F}f$, pour $f \in L^1(\mathbf{R})$, A(R) est donc un sous-espace de l'espace vectoriel C,,(R) des fonctions continues sur R qui tendent vers 0 à l'infini. En fait, A(R) est strictement plus petit que C,(R).

b) Sifet g sont intégrables, leur produit de convolution $f * g$, défini par :

$$(7) \qquad (f * g)(x) = \int_{\mathbf{R}} f(x-t) g(t) \, dt$$

l'est également. On a alors la relation :

$$\mathcal{F}(f * g) = (\mathcal{F}f)(\mathcal{F}g),$$

de sorte que A(R) est un anneau de fonctions continues sur R, de même que $L^1(\mathbf{R})$ est un anneau pour la convolution. Cette circonstance permet d'appliquer à l'étude de A(R) et de L'(R) la théorie des algèbres normées (cf. algèbres NORMÉES), qui en est d'ailleurs issue en grande partie.

c) Si y et u sont des réels, φ une fonction, on définit les fonctions φ_y, φ^u, $\bar{\varphi}$, $\check{\varphi}$, $\tilde{\varphi}$ par :

$$\begin{gathered} \varphi_y(x) = \varphi(x-y) \\ \varphi^u(x) = e^{2i\pi ux}\varphi(x), \\ \bar{\varphi}(x) = \overline{\varphi(x)}, \\ \check{\varphi}(x) = \varphi(-x), \\ \tilde{\varphi}(x) = \overline{\varphi(-x)}. \end{gathered}$$

On a alors :

$$\begin{gathered} \mathcal{F}(f_y) = (\mathcal{F}f)^{-y}, \\ \mathcal{F}(f^u) = (\mathcal{F}f)_u, \\ \mathcal{F}(\bar{f}) = (\mathcal{F}f), \\ \mathcal{F}(\tilde{f}) = (\overline{\mathcal{F}f}), \\ \mathcal{F}(\check{f}) = (\mathcal{F}f). \end{gathered}$$

d) Si f est dérivable et si f et f' sont intégrables, on a :

$$(8) \qquad \mathcal{F}(f')(t) = 2\, i\pi t(\mathcal{F}f)(t).$$

e) Sifest intégrable, ainsi que son produit par x, alors $\mathcal{F}f$ est dérivable et on a :

$$(9) \qquad \mathcal{F}(xf(x)) = -\frac{1}{2i\pi}(\mathcal{F}f)'.$$

Il est intéressant de voir comment certaines propriétés des fonctions se traduisent sur leurs transformées de Fourier.

Par exemple, les relations (8) et (9) montrent que plus une fonction est régulière (dérivable), plus sa transformée de

Fourier tend rapidement vers 0 à l'infini. Inversement, plus f tend rapidement vers 0 à l'infini, plus $\mathcal{F}f$ est régulière. Voici un autre exemple, présenté en termes vagues : plus les valeurs d'une fonction sont concentrées autour de l'origine, plus celles de sa transformée de Fourier sont, au contraire, étalées.

Le théorème de réciprocité

De même que la série de Fourier d'une fonction périodique caractérise celle-ci, sans qu'aucune propriété de convergence soit nécessaire, la transformée de Fourier d'une fonctionfcaractérise cette fonction. Donc la donnée de $\mathcal{F}f$ contient toute l'information relative à f. Dans certains cas, il est possible d'exprimer f explicitement à partir de $\mathcal{F}f$.

Théorème. Si f et $\mathcal{F}f$ sont toutes deux intégrables, on a :

$$(10) \qquad f = \overline{\mathcal{F}}(\mathcal{F}f).$$

C'est là une remarquable propriété de symétrie entre les opérateurs $\mathcal{F}$ et $\overline{\mathcal{F}}$.

On peut encore interpréter cela comme une propriété de synthèse : si on appelle spectre defle support de sa transformée de Fourier $\mathcal{F}f$, c'est-à-dire l'adhérence de l'ensemble des points où $\mathcal{F}f$ ne s'annule pas, le théorème de réciprocité, lorsque $\mathcal{F}f$ est intégrable, s'écrit :

$$f(x) = \int_{\mathbf{R}} (\mathcal{F}f)(t) e^{2i\pi tx}\, dt$$

et exprime que f est, en un certain sens, synthétisable, puisque $f(x)$ s'exprime sous la forme d'une intégrale (donc comme limite de combinaisons linéaires) à partir des exponentielles qui correspondent à des valeurs t contenues dans son spectre.

Une classe très importante de fonctions se prête à l'utilisation du théorème de réciprocité : c'est la classe S des fonctions indéfiniment dérivables à décroissance rapide. Une fonction f appartient à S si, et seulement si, elle admet des dérivées de tous les ordres et si, quels que soient les entiers n et p :

$$\lim_{x \to \pm\infty} \left| x^n \frac{d^p f(x)}{dx^p} \right| = 0.$$

Les relations (8) et (9) montrent que si f appartient à S, il en est de même pour $\mathcal{F}f$, et réciproquement. De sorte que $\mathcal{F}$ est un isomorphisme linéaire de $\mathcal{S}$ sur S. dont $\overline{\mathcal{F}}$ est l'isomorphisme réciproque. En fait, si on considère la topologie usuelle de $\mathcal{S}$ (cf. DISTRIBUTIONS), $\mathcal{F}$ et $\overline{\mathcal{F}}$ sont des isomorphismes continus.

Extension aux distributions tempérées

Rappelons brièvement que, si l'on désigne par S' l'espace des distributions tempérées, espace dual de $\mathcal{S}$, la transformation de Fourier sur S permet de définir, par transposition, la notion de transformation de Fourier sur S'qui fournit un isomorphisme de S' sur S'. Cela donne d'intéressantes applications, par exemple, à l'étude d'équations différentielles. ou plus généralement d'équations de convolution (lorsque l'on étend cette théorie aux espaces $\mathbf{R}^n$ de dimension supérieure à 1, on peut traiter par ce procédé des équations aux dérivées partielles linéaires à coefficients constants). D'après la propriété b, une telle équation est transformée par Fourier en une équation qui pourra, dans certains cas, se résoudre par division. Par exemple, un opérateur différentiel linéaire à coefficients constants (convolution par une combinaison linéaire de dérivées de la mesure de Dirac en 0) devient, par transformation de Fourier, l'opérateur de multiplication par un polynôme.

Ces idées sont à rapprocher de la transformation de Laplace (cf. calcul SYMBOLIQUE).

Transformation de Fourier-Plancherel dans L^2 (R)

Les espaces L'(R) et L'(R) ne sont pas inclus l'un dans l'autre. Mais ils contiennent tous deux l'ensemble K(R) des fonctions continues nulles hors d'un ensemble borné, et tout élément de L'(R) ou de L*(R) peut être approché, au sens de L^1 ou de L^2 selon le cas, par des éléments de K(**R**).

Si f appartient à K(R), $\mathcal{F}f$ appartient à $L^2(\mathbf{R})$, et on a l'égalité :

$$(11) \qquad \int_{\mathbf{R}} |f(x)|^2 dx = \int_{\mathbf{R}} |\mathcal{F}f(t)|^2 dt,$$

formellement analogue à (4).

Cette relation permet d'étendre à L'(R) la transformation de Fourier (aussi bien **que** la transformation de Fourier réciproque) : on obtient ainsi un opérateur, toujours noté $\mathcal{F}$, de $L^2(\mathbf{R})$ dans $L^2(\mathbf{R})$ qui coïncide avec la notion précédente de transformation pour les fonctions qui, comme celles de K(R), appartiennent à la fois à L'(R) et à $L^2(\mathbf{R})$.

Pour f dans L'(R), on écrit aussi parfois :

$$(\mathcal{F}f)(t) = \int_{\mathbf{R}} f(x) e^{-2i\pi xt} dx,$$

bien que cette formule n'ait pas de sens pour deux raisons : d'abord, l'intégrale peut ne pas être convergente ; ensuite $\mathcal{F}f$ est un élément de L'(R), donc défini à un ensemble de mesure nulle près, et on ne peut pas considérer la valeur d'un tel objet en un point particulier.

Ce nouvel opérateur de $L^2(\mathbf{R})$ dans $L^2(\mathbf{R})$, appelé transformation de Fourier-Plancherel, est un isomorphisme isométrique de $L^2(\mathbf{R})$, dont l'isomorphisme réciproque est 3 (plus exactement, l'extension de $\overline{\mathcal{F}}$ à L^2).

Pour f et g dans $L^2(\mathbf{R})$, on a l'identité de Parseval, qui exprime précisément l'isomorphisme de l'espace de Hilbert L*(R) dans lui-même défini par $\mathcal{F}$:

$$\int_{\mathbf{R}} \bar{f}g = \int_{\mathbf{R}} \overline{(\mathcal{F}f)}(\mathcal{F}g).$$

Cela permet, comme pour les séries trigonométriques, d'exploiter, dans l'analyse harmonique des fonctions de carré intégrable, la puissance de la théorie des espaces de Hilbert.

Généralisations

La transformation de Fourier que nous avons introduite pour les fonctions d'une variable réelle se généralise sans peine aux fonctions de plusieurs variables réelles.

Soitfune fonction intégrable sur $\mathbf{R}^n$ (n entier $\geqslant 1$). On définit la transformée de Fourier de f comme la fonction de n variables définie sur $\mathbf{R}^n$ par :

$$(\mathcal{F}f)(t_1, t_2, ..., t_n) = \int_{\mathbf{R}^n} f(x_1, x_2, ..., x_n) \times \exp[-2i\pi(t_1x_1 + \quad + t_nx_n)]\, dx_1 \ldots dx_n.$$

Les propriétés sont, en dimension n, tout à fait analogues à celles que l'on a en dimension 1.

Une autre direction de généralisation consiste à définir la transformée de Fourier defnon seulement pour des valeurs réelles de t, mais aussi pour certaines valeurs complexes. On pose :

$$(\mathcal{F}f)(z) = \int_{\mathbf{R}} f(x) e^{-2i\pi zx} dx,$$

fonction définie pour certaines valeurs de z. Lorsquefest, par exemple, nulle pour les valeurs négatives de x, on retrouve la notion de transformée de Laplace (cf. calcul SYMBOLIQUE).

Lorsque $\mathcal{F}f(z)$ peut être définie pour toute valeur de z, on peut déduire du

comportement de la fonction de variable complexe obtenue de nombreux renseignements sur la fonction f'donnée (théorème de Paley-Wiener, par exemple).

4. Les groupes commutatifs localement compacts

La mesure de Haar

La démonstration par Haar, en 1933, de l'existence d'une mesure invariante par translation, sur une large classe de groupes topologiques, permet, à partir de cette époque, de situer l'analyse harmonique dans sa vraie perspective et d'en comprendre la nature profonde.

Si on considère, sur R, la mesure de Lebesgue dx, on constate qu'elle est invariante par translation, en ce sens que, pour toute fonction intégrable f et tout réel t, la translatée f_t est intégrable et a même intégrale que f. De même, sur le groupe multiplicatif des nombres complexes de module 1, que l'on peut, en ce qui concerne la théorie de la mesure, identifier à l'intervalle $[0, 2\pi]$, la mesure de Lebesgue est invariante par translation, car, pour toute fonction intégrable f, et tout A de module 1, on a :

$$\int_0^{2\pi} f(\lambda e^{it})\,dt = \int_0^{2\pi} f(e^{it})\,dt.$$

Parmi les groupes topologiques, ceux qui sont ***localement compacts*** (cette classe contient, entre autres, les groupes de Lie) possèdent une propriété analogue.

Pour une fonction f définie sur un groupe G, et un élément t du groupe, on considère les translatées ${}_tf$ et f_t de f par t, respectivement à gauche et à droite, données par :

$${}_tf(x) = f(t^{-1}x),\ f_t(x) = f(xt^{-1})$$

(le groupe est noté multiplicativement). Il y a lieu de distinguer les translations à gauche et à droite si G n'est pas commutatif. Lorsque G possède la propriété de compacité locale, le théorème de Haar affirme l'existence (et l'unicité à un facteur multiplicatif près) d'une mesure (cf. INTÉGRATION ET MESURE) invariante par les translations à gauche, c'est-à-dire telle que, pour tout $t \in$ G et toute fonction intégrable f, ${}_tf$ soit intégrable et de même intégrale que f. Il y a également une mesure invariante par les translations à droite, mais elle diffère en général de la mesure invariante à gauche (pour les groupes compacts, ces deux types de mesures sont identiques, de même que pour les groupes commutatifs). Une telle mesure est appelée *mesure* ***de*** *Haar* à ***gauche*** (ou à droite, selon le cas). Si μ est une mesure de Haar à gauche, f une fonction μ-intégrable et t un élément de G, on a donc :

$$\int {}_tf\,d\mu = \int f\,d\mu.$$

Le théorème de Haar et l'étude des représentations linéaires des groupes topologiques forment le cadre de l'analyse harmonique abstraite.

Nous allons donner un bref aperçu de cette théorie dans le cadre des groupes commutatifs localement compacts, où elle est beaucoup plus simple et plus développée.

Le théorème de dualité de Pontriaguine et Van Kampen

Soit G un groupe commutatif localement compact ; l'opération de G est notée additivement, 0 désigne l'élément neutre.

On appelle *caractère* de G tout homomorphisme continu de G dans le groupe multiplicatif des nombres complexes de module 1. Autrement dit, un caractère est

une fonction continue y sur G, telle que, quels que soient x et y dans G :

$$|\gamma(x)| = 1,$$
$$\gamma(x-y) = \frac{\gamma(x)}{\gamma(y)} = \gamma(x)\gamma(y);$$

on en déduit, évidemment :

$$\gamma(0) = 1,\ \gamma(-x) = \gamma(x),$$
$$\gamma(x+y) = \gamma(x)\gamma(y).$$

Si y et y' sont deux caractères, la fonction qui, à tout x de G, associe $\gamma(x)\gamma'(x)$ est encore un caractère que l'on note y + y'. 11 est facile de voir que l'on peut ainsi munir l'ensemble des caractères sur G d'une structure de groupe commutatif, que l'on rend localement compact en y considérant une topologie particulière (c'est la topologie de la convergence uniforme sur les parties compactes de G).

Notons $\hat{G}$ le groupe commutatif localement compact formé des caractères de G. Ce groupe est appelé *dual* de G. On peut considérer le dual de $\hat{G}$, qui est un groupe commutatif localement compact $\hat{\hat{G}}$.

Pour tout x de G, la fonction définie sur $\hat{G}$ par y → $\gamma(x)$ est un caractère sur G, donc un élément de $\hat{\hat{G}}$. On a ainsi une application de G dans $\hat{\hat{G}}$.

Le résultat fondamental établi, vers 1935, par Pontriaguine et Van Kampen est le suivant : L'application de G dans $\hat{\hat{G}}$ définie ci-dessus est un isomorphisme topologique. En d'autres termes, le dual du dual d'un groupe G s'identifie à G lui-même.

Eu égard à la dualité, on notera indifféremment :

$$\gamma(x),\ x(\gamma),\ <\gamma, x>,\ <x,\gamma>$$

le nombre complexe égal à la valeur en $x \in$ G du caractère y sur G, ou à la valeur en y $\in \hat{G}$ du caractère x sur $\hat{G}$.

Soit par exemple, Z le groupe additif des entiers relatifs. Un caractère y sur Z est déterminé par la valeur $\gamma(1) = \alpha$, car y(n) = α^n pour tout $n \in$ Z. Réciproquement, tout nombre *a* de module 1 définit un caractère y sur Z (tel que y(1) = a), par la formule y(n) = a". On identifie ainsi le dual de Z au groupe multiplicatif T des nombres complexes de module 1. D'après le théorème de dualité, les caractères sur T correspondent aux nombres entiers : tout caractère sur T est de forme $a \rightarrow \alpha^n$ pour un entier n ; cela peut d'ailleurs se voir directement. Pour le groupe R des nombres réels, un caractère est une fonction continue φ telle que :

$$|\varphi(x)| = 1,\ \varphi(x-y) = \varphi(x)\varphi(y);$$

une telle fonction est nécessairement de la forme x $\longrightarrow e^{2i\pi tx}$, où t est un paramètre réel ; à la somme t + t' correspond le produit des caractères définis par t et t' ; ainsi le groupe dual de R s'identifie à R lui-même.

La transformation de Fourier

Soit G un groupe commutatif localement compact, $\hat{G}$ le groupe dual de G, dx une mesure de Haar sur G.

À toute fonction intégrablef sur G, on associe une fonction $\hat{f}$ sur $\hat{G}$, la *transformée de Fourier de f*, définie, pour tout y $\in \hat{G}$, par :

$$\hat{f}(\gamma) = \int_G f(x) <\gamma, x> dx.$$

Si G = T, $\hat{G}$ = Z et la fonctionf est alors une suite, qui n'est autre que la suite des coefficients de Fourier de f (on peut considérer f indifféremment comme une fonction sur T, ou comme une fonction sur R de période 2π : on identifie les points de R congrus à t modulo 2π au nombre complexe e^{it} de module 1) ; si G = R, $\hat{G}$ = R etf est la transformée de Fourier usuelle (sur R) de f. Ainsi l'étude des séries

et intégrales de Fourier apparaît comme un cas particulier de la transformation de Fourier abstraite.

La transformation de Fourier abstraite jouit de propriétés semblables à celles que nous avons vues dans le cas des fonctions de variable réelle. Citons-en quelques-unes.

a) Si f est intégrable sur G, $\hat{f}$ est continue sur $\hat{G}$.

b) $f \equiv 0$ si, et seulement si, $\hat{f} \equiv 0$. Autrement dit, $\hat{f}$ caractérise parfaitement f.

c) Soit f et g deux fonctions intégrables sur G. Alors la fonction $f * g$, définie sur G par :

$$(f*g)(x) = \int_G f(x-y)g(y)\,dy$$

est intégrable, et : $\widehat{(f*g)} = \hat{f}\,\hat{g}$.
Ainsi, l'ensemble A(G) des transformées de Fourier des fonctions intégrables sur G est une algèbre. Les méthodes générales de la théorie des algèbres normées s'appliquent à l'étude de nombreuses propriétés de A(G), ou aussi bien de l'algèbre de convolution $L^1(G)$.

d) *Formule de réciprocité.* On peut choisir les mesures de Haar dx et $d\gamma$ de G et de $\hat{G}$ (qui dépendent d'un facteur constant que l'on peut ajuster) de telle sorte que, lorsque f est intégrable sur G, et $\hat{f}$ intégrable sur $\hat{G}$, l'on ait, pour tout $x \in G$ et tout $\gamma \in \hat{G}$:

$$\hat{f}(\gamma) = \int_G f(x)\,\overline{<\gamma, x>}\,dx,$$

$$f(x) = \int_{\hat{G}} \hat{f}(\gamma) <\gamma, x>\,d\gamma.$$

e) *Théorème de Bessel-Parseval-Plancherel* : Si f est à la fois intégrable et de carré intégrable sur G, $\hat{f}$ est de carré intégrable sur $\hat{G}$ et on a, avec le même choix des mesures de Haar que ci-dessus :

$$\int_G |f(x)|^2\,dx = \int_{\hat{G}} |\hat{f}(\gamma)|^2\,d\gamma.$$

Cela permet, comme dans le cas réel, de définir une transformation de Fourier-Plancherel qui est un isomorphisme isométrique de $L^2(G)$ sur $L^2(\hat{G})$. Ici encore, les méthodes propres aux espaces de Hilbert s'appliquent avec fruit.

Ainsi, l'étude des groupes commutatifs localement compacts offre à l'analyse harmonique son cadre naturel, et l'étude abstraite permet de retrouver et d'approfondir la théorie de l'analyse harmonique sur R et celle des séries de Fourier. Bien entendu, des propriétés liées à la structure particulière de R (comme celles qui se rattachent à la dérivation, l'extension aux distributions, par exemple) ne sont pas susceptibles d'un tel traitement abstrait. Aussi l'analyse harmonique continue-t-elle à se développer sur les deux plans, dans le cadre abstrait des groupes commutatifs localement compacts, et dans le cadre classique de la droite réelle et des espaces $\mathbf{R}^n$.

Notons enfin que, si l'on considère des groupes localement compacts non commutatifs, il existe – surtout dans le cadre des groupes de Lie – une théorie similaire, mais beaucoup plus éloignée de la situation « classique » de l'analyse harmonique sur R (par exemple l'objet dual n'est plus un groupe, la transformée de Fourier d'une fonction n'est plus une fonction numérique, etc.), et qui fait actuellement l'objet d'un développement rapide.

RENÉ SPECTOR

Bibliographie

R. N. BRACEWELL, *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2ᵉ éd. 1986 / J.-L. CLERC, P. EYMARD, J. FARAUT et al., *Analyse harmonique*, Centre international de mathématiques pures et appliquées, Nice, 1983 / M. HERVÉ, *Transformation de Fourier et distributions*, P.U.F., Paris, 1986, 2 vol., Springer-Verlag, New York.

1987-1988 / J.-P. PIER, *L'Analyse harmonique : son développement historique*. Masson. Paris, 1990 / M. SAMUELIDES & L. TOUZILLIER, *Analyse harmonique*, Cépadues. Toulouse, 1990 / V.S. VARADARAJAN, *An Introduction to Harmonic Analysis on Semisimple Lie Groups*, Cambridge Univ. Press, 1989 / A. ZYGMUND, *Trigonometrical Series*, 2 vol., *ibid.* 1988.

HARMONIQUES FONCTIONS

→ POTENTIEL & FONCTIONS HARMONIQUES

HILBERT ESPACE DE

La théorie des espaces hilbertiens trouve son origine dans celle des développements de fonctions arbitraires en séries de fonctions orthogonales, lesquelles apparaissent le plus souvent comme fonctions propres de certains opérateurs différentiels linéaires (séries de Fourier, fonctions sphériques, théorie des oscillations de Sturm-Liouville). À l'occasion de l'étude des équations intégrales, ébauchées par V. Volterra, I. Fredholm et E. Schmidt, Hilbert définit l'espace l^2 des suites de carré sommable, et résout les principaux problèmes posés en interprétant les équations en termes d'endomorphismes de l'espace l^2. E. Schmidt. M. Fréchet et F. Riesz donnent ensuite une forme plus géométrique à la théorie de Hilbert, en introduisant le langage des normes, de l'orthogonalité et des bases hilbertiennes, et découvrent que de nombreux espaces fonctionnels classiques sont isomorphes à l^2, ou à des sous-espaces vectoriels de cet espace. Dès lors s'impose une présentation axiomatique des espaces préhilbertiens et hilbertiens ; elle est essentiellement due à J. von Neumann et à F. Riesz ; le lecteur la trouvera esquissée ci-dessous. Enfin, ces derniers approfondissent considérablement l'étude des endomorphismes des espaces hilbertiens, et créent ainsi un des outils les plus puissants de l'analyse fonctionnelle et de la physique mathématique.

Nous supposons connus les notions fondamentales de l'algèbre linéaire (cf. algèbre LINÉAIRE ET MULTILINÉAIRE), le langage des normes et semi-normes, et la notion de famille sommable (cf. SÉRIES ET PRODUITS INFINIS).



Généralités

Espaces préhilbertiens

On appelle espace vectoriel *préhilbertien* (complexe) un espace vectoriel sur le corps C des nombres complexes, muni d'une forme sesquilinéaire autoadjointe dont la forme hermitienne associée est positive, c'est-à-dire d'une application de E X E dans C, notée $(x, y) \mapsto (x|y)$, satisfaisant aux conditions suivantes :

- pour **tout** élément y de E, l'application $x \mapsto (x|y)$ est linéaire ;
- pour tout couple (x, y) d'éléments de E, $(y|x) = \overline{(x|y)}$;

pour tout élément x de E, $(x|x) \geqslant 0$.

Le scalaire $(x|y)$ s'appelle *produit hermitien* des vecteurs x et y.

On dit que l'espace vectoriel E est préhilbertien séparé, *ou hermitien,* **si la** forme hermitienne considérée est définie positive, c'est-à-dire si la relation $(x|x) = 0$ implique la relation $x = 0$.

Théorème 1. Pour tout couple (x, y) d'éléments d'un espace préhilbertien E :

$$|(x \mid y)|^2 \leqslant (x \mid x) . (y \mid y)$$

(inégalité de Schwarz).

Écartons le cas où l'un des deux vecteurs x et y est nul. Écrivons que, pour tout nombre réel α et pour tout nombre complexe β de module 1, le nombre réel :

$$\begin{aligned}(x + \alpha\beta y \mid x + \alpha\beta y) &= \alpha\beta\alpha\beta(y \mid y) + \alpha\beta(x \mid y) \\ &\quad + \alpha\beta(y \mid x) + (x \mid x) \geqslant 0,\end{aligned}$$

ou encore :

$$\alpha^2(y \mid y) + 2\alpha \operatorname{Re}[\bar{\beta}(x \mid y)] + (x \mid x) \geqslant 0 ;$$

par suite, le discriminant de ce trinôme du second degré en α est négatif ou nul pour tout nombre complexe β de module 1,

$$(\operatorname{Re}[\beta(x \mid y)])^2 \leqslant (x \mid x) . (y \mid y).$$

L'inégalité cherchée étant évidente lorsque $(x \mid y) = 0$, écartons ce cas. Nous obtenons alors l'inégalité de Schwarz en posant :

$$\beta = \frac{(x \mid y)}{|(x \mid y)|}.$$

Lorsque l'espace vectoriel E est hermitien, on montre qu'il y a égalité dans l'inégalité de Schwarz si et seulement si les vecteurs x et y sont colinéaires.

Théorème 2. Soit E un espace vectoriel préhilbertien. L'application qui à tout vecteur x de E associe le nombre réel positif $\|x\| = (x \mid x)^{1/2}$ est une semi-norme sur E, dite associée à la forme sesquilinéaire $(x, y) \mapsto (x \mid y)$.

En effet, pout tout nombre complexe α, $\|\alpha x\| = |\alpha| . \|x\|$. Pour tout couple (x, y) de vecteurs de E :

$$(1) \quad \|x + y\|^2 = (x + y \mid x + y) = \|x\|^2 + \|y\|^2 + 2 \operatorname{Re}(x \mid y).$$

D'autre part :

$$(2) \quad (\|x\| + \|y\|)^2 = \|x\|^2 + \|y\|^2 + 2 \|x\| . \|y\|.$$

L'inégalité triangulaire :

$$\|x + y\| \leqslant \|x\| + \|y\|$$

découle des relations (1) et (2), de la relation $\operatorname{Re}(x \mid y) \leqslant |(x \mid y)|$ et de l'inégalité de Schwarz.

La semi-norme précédente est une norme si et seulement si l'espace vectoriel E est hermitien. Le nombre réel positif $\|x\|$ s'appelle alors norme hermitienne du vecteur x, et le nombre $\|x - y\|$ distance hermitienne des points x et y. Un vecteur de norme 1 est dit unitaire. Dans ces conditions, il y a égalité dans l'inégalité triangulaire si et seulement si les vecteurs x et y sont colinéaires et de même sens, c'est-à-dire s'il existe un couple (α, β) de nombre réels positifs non tous deux nuls tel que $\alpha x = \beta y$.

Espaces hilbertiens

En algèbre, on utilise surtout les espaces hermitiens de dimension finie. En analyse, ce sont les espaces hermitiens de dimension infinie qui interviennent dans la plupart des questions ; on est amené à supposer que ces espaces sont complets, c'est-à-dire que toute suite de Cauchy est convergente. Un espace hermitien complet est dit *hilbertien.* Tout espace hermitien de dimension finie est hilbertien.

Voici deux exemples fondamentaux :

Soit I un ensemble non vide. L'espace vectoriel $\mathbf{C}^{(\mathrm{I})}$ des applications de I dans C nulles sauf pour un nombre *fini* de valeurs de la variable, muni de la forme sesquilinéaire qui, aux vecteurs $x = (\xi_i)$ et $y = (\eta_i)$, $i \in \mathrm{I}$, associe le nombre complexe :

$$\sum_{i \in \mathrm{I}} \xi_i \overline{\eta_i},$$

est hermitien ; il est hilbertien si et seulement si l'ensemble 1 est fini.

Soit I?(I) l'espace vectoriel des familles $x = (\xi_i)$, $i \in 1$, de nombres complexes telles que :

$$\sum_{i \in I} |\xi_i|^2 < +\infty.$$

Si $x = (\xi_i)$ et y $= (\eta_i)$, $i \in 1$, sont deux éléments de cet espace, la famille $(\xi_i\overline{\eta_i})$, $i \in I$, est sommable. Muni de l'application :

$$(x, y) \mapsto \sum_{i \in I} \xi_i \overline{\eta_i},$$

l^2 (I) est un espace hilbertien.

Voici deux autres exemples, dont l'importance est capitale en analyse fonctionnelle. Soit I un intervalle de R non réduit à un point, et p une fonction à valeurs réelles continue sur I, prenant des valeurs strictement positives en tout point intérieur a I. Soit $\mathcal{C}(I, p)$ l'espace vectoriel des fonctions f continues sur I à valeurs complexes telles que :

$$\int_I |f(t)|^2 p(t)\, dt < +\infty.$$

Pour tout couple (f, g) d'éléments de cet espace vectoriel. fgp est intégrable sur I. Muni de l'application :

$$(3) \qquad (f, g) \mapsto \int_I f(t)\overline{g(t)}\, p(t)\, dt,$$

$\mathcal{C}(I, p)$ est un espace vectoriel hermitien, mais ce n'est pas un espace hilbertien. En revanche, l'espace vectoriel $L^2(I, p)$ des classes de fonctionsfà valeurs complexes, mesurables sur I et telles que :

$$\int_I |f(t)|^2 p(t)\, dt < +\infty,$$

muni de l'application définie par la formule (3), est hilbertien, ce qui met en évidence l'intérêt de la théorie de Lebesgue pour toutes ces questions.

Orthogonalité

On dit que deux vecteurs x et y d'un espace hermitien E sont orthogonaux si leur produit hermitien est nul : $(x|y) = 0$. Puisque $(y \; x) = (x \; y)$, cette relation est symétrique.

On dit que deux parties A et B de E sont orthogonales si, pour tout élément x de A et pour tout élément y de B, (xl y) = 0. L'ensemble, noté $A^\perp$, des vecteurs orthogonaux à une partie A de E est un sous-espace vectoriel fermé de E, appelé orthogonal de A. L'orthogonal de E est réduit au vecteur nul.

Soit F un sous-espace vectoriel de E. On dit que F admet un supplémentdire orthogonal s'il existe un sous-espace vectoriel G de E orthogonal à F tel que $E = F \oplus G$. Alors $G = F^\perp$; c'est pourquoi $F^\perp$ s'appelle le supplémentaire orthogonal de F dans E. La projection d'un vecteur x de E sur F parallèlement à G s'appelle projection orthogonale de x sur F. Sous ces mêmes hypothèses, $(F^\perp)^\perp = F$. En effet, il est évident que $(F^\perp)^\perp$ contient F. Soit donc x un élément de $(F^\perp)^\perp$; écrivons x sous la forme $x = y + z$, où $y \in F^\perp$ et $z \in F$. Il s'ensuit que :

$$(x|y) = (y|y) + (z|y),$$

c'est-à-dire que $(y|y) = 0$. Ainsi, $y = 0$, et $x = z$. Le vecteur x appartient donc à F.

Soit maintenant (F_i), $i \in 1$, une famille de sous-espaces vectoriels de E orthogonaux deux à deux, et F leur somme. Alors cette somme est directe ; c'est pourquoi l'on dit que F est somme directe orthogonale des sous-espaces vectoriels F,.

Soit $S = (x_i)$, $i \in I$, une famille de vecteurs d'un espace hermitien E. On dit que S est orthogonale si, pour tout couple

(i, j) d'éléments distincts de I, les vecteurs x_i et x_j sont orthogonaux. Dans ces conditions, pour toute partie finie J de I :

$$\|\sum_{i \in J} x_i\|^2 = \sum_{i \in J} \|x_i\|^2$$

(théorème de Pythagore). On dit que S est orthonormale si, de plus, pour tout élément i de 1, le vecteur x_i est unitaire ; la somme des droites $\mathbf{C}x_i$ est alors directe orthogonale.

On dit enfin que S est une base *hilbertienne* de E si S est orthonormale, et si le sous-espace vectoriel engendré par S est dense dans E. Cette notion est mieux adaptée à l'analyse que celle de base orthonormale.

Par exemple, la base canonique (e_i), $i \in 1$, de l'espace vectoriel $\mathbf{C}^{(I)}$ est orthonormale ; cette famille est une base hilbertienne, dite canonique, de l'espace hilbertien $l^2(\mathrm{I})$, car $\mathbf{C}^{(I)}$ est dense dans $l^2(\mathrm{I})$.

Voici un autre exemple, lié de manière essentielle à la théorie des séries de Fourier. Soit $\mathcal{C}(\mathbf{T})$ l'espace vectoriel des fonctions continues sur R à valeurs complexes et admettant 1 pour période, muni du produit hermitien :

$$(f, g) \mapsto \int_0^1 f(t)\overline{g(t)}\, dt.$$

La famille (e_n), $n \in \mathbf{Z}$, des fonctions définies par les formules :

$$e_n(t) = e^{2i\pi nt}$$

est une base hilbertienne de C(T) ; les éléments du sous-espace vectoriel engendré par S s'appellent polynômes trigonométriques. L'espace hermitien C(T) n'est pas complet; il peut s'identifier à un sous-espace vectoriel de l'espace vectoriel hilbertien $L^2([0, 1])$, dense dans $L^2([0, 1])$. La famille (e_n), $n \in \mathbf{Z}$, apparaît alors comme une base hilbertienne de $L^2([0, 1])$.

Théorie élémentaire

L'étude des espaces hermitiens de dimension finie repose sur le théorème qui suit.

Théorème 3. Tout espace hermitien de dimension finie admet au moins une base orthonormale.

La démonstration s'effectue par récurrence sur la dimension de l'espace hermitien E. Soit donc E un espace hermitien de dimension strictement positive n. Choisissons un vecteur unitaire e_1. L'ensemble H des vecteurs orthogonaux à e_1 est un hyperplan de E, car c'est le noyau de la forme linéaire non nulle $x \mapsto (x \mid e_1)$. De plus, e_1 n'appartient pas à H, si bien que E est somme directe orthogonale de la droite Ce_1 et de H. Il suffit alors d'appliquer l'hypothèse de récurrence à H, qui est de dimension $n - 1$, pour obtenir une base orthonormale de E.

Théorème 4. Soit E un espace hermitien et F un sous-espace vectoriel de E de dimension finie.

- Pour tout vecteur x de E, il existe un couple (y, z) et un seul de vecteur de E tel que $y \in F^{\perp}$, $z \in F$ et $x = y + z$. Autrement dit, le sous-espace vectoriel $F^{\perp}$ est supplémentaire orthogonal de F dans E. Par suite, $(F^{\perp})^{\perp} = F$. Enfin, pour tout vecteur u de F différent de z, $\|x - u\| > \|x - z\|$.
- Si F est muni d'une base orthonormale $(e_1, e_2, \ldots, e_p)$, alors :

$$z = \sum_{j=1}^{p} (x \mid e_j) e_j.$$

Munissons F d'une base orthonormale $(e_1, e_2, ..., e_p)$. Le vecteur z, s'il existe, peut s'écrire d'une manière et d'une seule sous la forme $z = \alpha_1 e_1 + \dots + \alpha_p e_p$; or.

$$(x \mid e_j) = (y \mid e_j) + (z \mid e_j) = \alpha_j ;$$

par suite, les vecteurs y et z sont nécessairement définis par les formules :

$$z = \sum_{j=1}^{p} (x \quad e_j)e_j \quad \text{et} \quad y = x - z,$$

ce qui prouve l'unicité du couple (y, z). Réciproquement, les vecteurs ainsi définis conviennent visiblement.

Soit maintenant u un vecteur de F. Le vecteur $x - z$ est orthogonal à F, et le vecteur $z - u$ appartient à F. Donc :

$$\|x - u\|^2 = \|x - z\|^2 + \|z - u\|^2,$$

ce qui achève la démonstration.

Corollaire 1. Toute famille orthonormale d'éléments d'un espace hermitien E de dimension finie peut être complétée en une base orthonormale de E.

Soit en effet F le sous-espace vectoriel engendré par une famille orthonormale L. D'après le théorème, $F^{\perp}$ est supplémentaire orthogonal de F dans E. Il existe une base orthonormale L' de $F^{\perp}$. La base de E obtenue en réunissant L et L' convient.

Corollaire 2. Soit $(e_1, e_2, \ldots, e_p)$ une famille orthonormale de vecteurs d'un espace hermitien E, et x un vecteur de E. Pour tout élément j de $[1, n]$, on pose $\xi_j = (x \quad e_j)$. Alors la fonction f qui à tout élément (A,, A,, ... A,,) de $\mathbf{C}^p$ associe le nombre réel positif :

$$f(\lambda_1, \lambda_2, \ldots, \lambda_p) = \|x - \sum_{j=1}^{p} \lambda_j e_j\|$$

admet un minimum strict au point $(\xi_1, \xi_2, \ldots, \xi_p)$. Autrement dit, pour approcher le mieux possible (en norme) un vecteur x par des éléments de la forme :

$$\sum_{j=1}^{p} \lambda_j e_j,$$

il convient de prendre $\lambda_j = \xi_j$ pour tout $j \in [1, p]$. Posons en effet :

$$z = \sum_{j=1}^{p} \xi_j e_j \quad \text{et} \quad u = \sum_{j=1}^{p} \lambda_j e_j$$

Puisque u appartient au sous-espace vectoriel F engendré par $e_1, e_2, \ldots, e_p$, nous savons que, si $u \neq z$:

$$\|x - u\| > \|x - z\|,$$

ce qu'il fallait prouver.

Soit, par exemple, f une fonction continue à valeurs complexes admettant 1 pour période, et p un entier naturel. Parmi les polynômes trigonométriques de la forme :

$$\sum_{n=-p}^{p} \lambda_n e_n,$$

celui qui approche le mieuxfen moyenne quadratique est le polynôme :

$$S_p = \sum_{n=-p}^{p} (f \quad e_n)e_n,$$

somme partielle à l'ordre p de la *série de Fourier* de f.

Procédé d'orthonormalisation de Schmidt. Soit E un espace hermitien, $(e_1, e_2, \ldots, e_p)$ une famille orthonormale de vecteurs de E, et F le sous-espace vectoriel de E engendré par cette famille. On suppose que F est différent de E, et on considère un vecteur x de E n'appartenant pas à F. Il existe alors un vecteur e_{p+1} de E et un seul tel que :

- la famille $(e_1, e_2, \ldots, e_p, e_{p+1})$ soit orthonormale ;

le vecteur e_{p+1} appartienne au sous-espace vectoriel $F \oplus \mathbf{C}x$;

le scalaire $(e_{p+1} \quad x)$ soit réel positif.

De plus, le vecteur e_{p+1} est donné par la formule :

$$e_{p+1} = \frac{y}{\|y\|},$$

où :

$$y = x - \sum_{j=1}^{p} (x \mid e_j) e_j$$

Par récurrence, on en déduit le théorème suivant.

Théorème 5. Soit (x_n), $n \in N$, une famille libre d'éléments d'un espace hermitien E. Il existe alors une famille orthonormale (e_n), $n \in N$, et une seule de vecteurs de E telle que, pour tout entier n, e_n appartienne au sous-espace vectoriel engendré par les vecteurs $x_0, x_1, \ldots, x_n$, et que $(x_n \mid e_n)$ soit réel positif. On dit que (e_n), $n \in$ N, se déduit de (x_n), $n \in$ N, par orthonormalisation. Ces deux familles engendrent le même sous-espace vectoriel F de E.

En particulier, si (x_n), $n \in$ N, est totale, c'est-à-dire si F est dense dans E, (e_n), $n \in$ N, est une base hilbertienne de E. Il en résulte que tout espace hermitien séparable (c'est-à-dire, admettant une famille de vecteurs totale dénombrable) admet une base hilbertienne dénombrable.

Appliquons ces résultats à l'espace hermitien C(I, p) introduit plus haut, en supposant que, pour tout entier naturel n, la fonction $x_n : t \mapsto t^n$ est un élément de C(I, p). La famille (e_n), déduite de (x_n), $n \in$ N, par orthonormalisation est constituée de fonctions polynomiales, e_n étant de degré n. La famille (e_n) s'appelle système de polynômes orthogonaux associé au poids p sur l'intervalle I. Lorsque l'intervalle I est borné, (e_n), $n \in$ N, est une base hilbertienne de $\mathcal{C}(I, p)$; il en est de même lorsque I est non borné, s'il existe deux nombres réels strictement positifs α et β tels que, pour tout $t \in$ I, $p(t) \leqslant \beta \exp(-\alpha \mid t \mid)$ (cf. polynômes ORTHOGONAUX).

Étudions enfin les principales propriétés des bases hilbertiennes.

Théorème 6. Soit (e_i), $i \in$ I, une famille orthonormale d'éléments d'un espace hermitien E, F l'adhérence du sous-espace vectoriel engendré par les vecteurs e_i, x un vecteur de E, et (ξ_i), $i \in$ I, la famille des composantes de x suivant (e_i), $i \in$ I, c'est-à-dire des scalaires $\xi_i = (x \mid e_i)$.

1. La famille (ξ_i), $i \in$ I, est de carré sommable, et :

$$\sum_{i \in I} |\xi_i|^2 \leqslant \|x\|^2$$

(inégalité de Bessel).

2. Pour que x appartienne à F, il faut et il suffit que :

$$\sum_{i \in I} |\xi_i|^2 = \|x\|^2$$

(égalité de Parseval).

Dans ces conditions, la famille $(\xi_i e_i)$, $i \in$ I, est sommable, et :

$$x = \sum_{i \in I} \xi_i e_i.$$

3. Pour tout couple (x, y) d'éléments de F de composantes respectives (ξ_i) et (η_i), $i \in$ I, la famille $(\xi_i \overline{\eta_i})$, $i \in$ I, est sommable, et :

$$(x \mid y) = \sum_{i \in I} \xi_i \overline{\eta_i}.$$

Ce théorème est une conséquence immédiate du théorème 4, puisque, pour toute partie finie J de I, la projection orthogonale de x sur le sous-espace vectoriel F, engendré par la famille (e_i), $i \in$ J, est égale à :

$$\sum_{i \in J} \xi_i e_i.$$

Lorsque (e_i), $i \in I$, est une base hilbertienne de E, les assertions 2 et 3 s'appliquent à tous les éléments de E.

Corollaire. Soit E un espace hermitien et (e_i), $i \in I$, une base hilbertienne de E. Alors l'application qui à tout vecteur de E associe la famille de ses composantes dans la base (e,) est un isomorphisme de l'espace hermitien E sur un sous-espace vectoriel de l'espace hilbertien $l^2(I)$.

Si l'espace E est hilbertien, cette application est un isomorphisme de E sur $l^2(I)$.

En effet, si E est complet, pour **tout** élément (α_i), $i \in I$, de $1^2(I)$, la famille $(\alpha_i e_i)$, $i \in I$, est sommable. Alors le vecteur :

$$y = \sum_{i \in I} \alpha_i e_i$$

admet α_i pour i-ième composante.

On peut appliquer ce théorème aux développements en série de fonctions orthogonales (séries de Fourier, polynômes orthogonaux, etc.).

Espaces hilbertiens

Dans la théorie précédente, le théorème de projection orthogonale (théorème 4) a joué un rôle fondamental, Il ne s'étend malheureusement pas au cas d'un sous-espace vectoriel fermé quelconque F d'un espace hermitien. Ainsi, dans l'espace hermitien $\mathcal{C}([-1, 1])$, l'hyperplan fermé noyau de la forme linéaire continue :

$$f \mapsto \int_0^1 f(t)\, dt$$

n'admet pas de supplémentaire orthogonal, Néanmoins, si F est complet, le théorème 4 s'étend de la manière suivante :

Théorème 7. Soit E un espace hermitien et F un sous-espace vectoriel complet de E. Alors F admet un supplémentaire orthogonal, et $(F^{\perp})^{\perp} = F$.

Ce théorème contient le théorème 4 comme cas particulier, et s'applique aussi au cas où E est hilbertien et F fermé.

La démonstration s'appuie sur le théorème suivant.

Théorème 8. Soit E un espace hermitien, F une partie convexe complète non vide de E, et x un élément de E. Il existe alors un élément z de F et un seul tel que :

$$\|x - z\| = d(x, F),$$

où :

$$d(x, F) = \inf_{u \in F} \|x - u\|.$$

On montre pour cela que **toute** suite (z_n) de points de F telle que $\|x - z_n\|$ converge vers $d(x, F)$ est une suite de Cauchy. Comme F est complet, la suite (z_n) admet une limite z dans F, et on vérifie que z convient.

On en déduit facilement le théorème 7. en prouvant que $y = x - z$ est orthogonal à F.

Dégageons quelques conséquences du théorème 7.

Corollaire 1. Soit F un sous-espace vectoriel fermé d'un espace hilbertien E. Si $F \neq E$, il existe un vecteur non nul de E orthogonal à F.

Corollaire 2. Pour qu'une famille (e_i), $i \in I$, de vecteurs d'un espace hilbertien E soit totale, il faut et il suffit que le vecteur nul soit le seul vecteur orthogonal à tous les vecteurs e,.

Corollaire 3. Soit a un vecteur d'un espace hilbertien E. L'application f_a : $x \mapsto (x \mid a)$ est une forme linéaire dont la norme est égale à celle de a. Autrement dit :

$$\sup_{\|x\| \leqslant 1} (x \mid a) = \|a\|.$$

Réciproquement, pour toute forme linéaire continue f sur E, il existe un

vecteur a de E et un seul tel quefsoit égale à l'application $x \mapsto (x| a)$. Ainsi, l'application $a \mapsto f_a$ est une application semi-linéaire bijective de E sur son dual topologique E*.

Corollaire 4. Toute famille orthonormale d'éléments d'un espace hilbertien E peut être complétée en une base hilbertienne de E. En particulier, tout espace hilbertien admet au moins une base hilbertienne (e_i), $i \in I$. L'application

$$(\alpha_i) \mapsto \sum_{i \in I} \alpha_i e_i$$

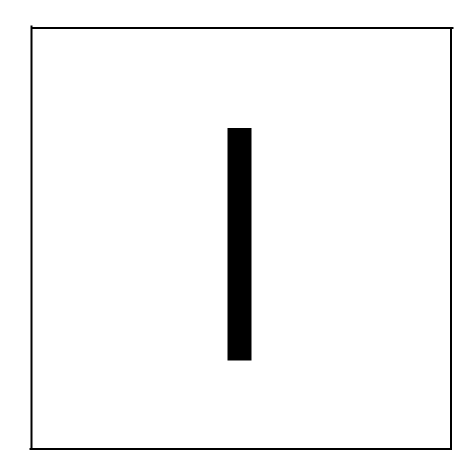
est alors un isomorphisme de l'espace hilbertien $\ell^2(I)$ sur E.

On démontre aussi que deux bases hilbertiennes d'un espace hilbertien E sont équipotentes. Le cardinal d'une base hilbertienne de E s'appelle dimension hilbertienne de E.

LUCIEN CHAMBADAL et JEAN-LOUIS OVAERT

Bibliographie

S. K. BERBERIAN, *Introduction to Hilbert Space*, Chelsea Publ., New York, 2' éd. 1991 / N. BOURBAKI, *Espaces vectoriels topologiques*, chap. V, Hermann, Paris, 1964, rééd.. Masson, 1981 / J. DIEUDONNÉ, *Éléments d'analyse*, Gauthier-Villars, Paris, t. 1, 2' éd. 1979, t. II, 4' éd. 1981 / J. DIXMIER, *Les Algèbres d'opérateurs dans l'espace hilbertien*, ibid.., 1969 / A. GUICHARDET, *Intégration : analyse hilbertienne*, éd. Marketing, Paris, 1989 / F. RIESZ & B. SZ. NAGY, *Leçons d'analyse fonctionnelle*, Gauthier-Villars. 1965, reprod. en fac-sim., J. Gabay, Sceaux, 1990 / L. SCHWARTZ, *Analyse hilbertienne*, Hermann, 1979 / J. WEIDMANN, *Linear Operators on Hilbert Spaces*, Springer, New York, 1980.



INTÉGRALES ÉQUATIONS

Les premières équations intégrales furent obtenues par Daniel Bernoulli vers 1730 dans l'étude des oscillations d'une corde tendue. Après l'introduction du noyau de Green, il fallut attendre les dernières années du XIXe siècle, avec les travaux de H. A. Schwarz, de H. Poincaré, de V. Volterra et surtout ceux de 1. Fredholm, pour disposer de résultats généraux en liaison étroite avec les premiers développements de l'analyse fonctionnelle. Quelques années plus tard, l'étude des équations intégrales conduisait D. Hilbert à définir l'espace qui porte son nom et à poser les premières bases de la théorie spectrale, cadre dans lequel F. Riesz développa la théorie des opérateurs compacts (19 18). Ainsi, les équations intégrales ont joué un rôle historique important dans l'élaboration des principaux concepts de l'analyse contemporaine.

E. U.

1. Exemples

La forme usuelle d'une équation intégrale est :

$$(1) \quad y(x) - \lambda \int_A K(x, \xi) y(\xi) \, d\xi = f(x),$$

où A est une partie de $\mathbf{R}^m$ décrite par chacune des variables x et ξ, K une fonction donnée sur A^2 appelée *noyau* de l'équation, f une fonction donnée sur A, qui est la constante 0 dans l'équation *homogène* :

$$(2) \quad y(x) = \lambda \int_A K(x, \xi) y(\xi) \, d\xi ;$$

enfin la fonction y est l'inconnue de l'équation et A un paramètre ; toutes ces quantités sont de préférence complexes.

Problème de Sturm-Liouville

Le problème de Sturm-Liouville (cf. équations DIFFÉRENTIELLES, chap. 3) concerne les valeurs du paramètre réel λ pour lesquelles l'équation différentielle linéaire homogène :

$$Ly - \lambda r y = 0$$

(où L est un opérateur différentiel d'ordre n à coefficients continus sur un intervalle compact [a. b] de R et r une fonction continue strictement positive sur cet intervalle) a des solutions non nulles vérifiant n conditions aux limites données.

Si 0 n'est pas l'une de ces valeurs de A, on définit une *fonction de Green G* du problème, continue sur [a, $b]^2$; si l'on connaît G, la formule intégrale :

$$y(x) = \int_a^b G(x, \xi) f(\xi) \, d\xi$$

donne la solution de l'équation non homogène :

$$Ly = f(x),$$

vérifiant les conditions aux limites données, de sorte que le problème de Sturm-Liouville est transformé en l'équation intégrale homogène :

$$y(x) = \lambda \int_a^b G(x, \xi) r(\xi) y(\xi) \, d\xi.$$

Problème de Dirichlet

Le problème de Dirichlet, dans un ouvert borné A de $\mathbf{R}^m$, pour une fonction continue fdonnée sur la frontière Γ de A, consiste à trouver la fonction, unique d'après le principe du maximum, continue sur :

$$\bar{\Delta} = \Delta \cup \Gamma,$$

harmonique sur A, qui coïncide avec f sur Γ. En 1877, C. G. Neumann proposait la méthode suivante pour la solution de ce problème, en supposant $m = 2$ et Γ pourvue d'une tangente continue ; on désignera par L(T) la longueur de la courbe Γ.

Soit (ξ, η) le point courant de Γ, d'abscisse curviligne σ, et (α, β) les cosinus directeurs de la normale en ce point orientée vers A. On appelle *potentiel de double couche* d'une densité continue μ sur Γ la limite, quand δ tend vers 0, du quotient par 2 δ de la différence entre le potentiel de la densité μ au point $(\xi + \alpha\delta, \eta + \beta\delta)$ et celui de la densité $-\mu$ au point (ξ, η). Le potentiel de double couche est la fonction :

$$h(x, y) = \int_0^{L(\Gamma)} \mu(\sigma) \frac{\alpha(x - \xi) + \beta(y - \eta)}{(x - \xi)^2 + (y - \eta)^2} \, d\sigma = \int_\Gamma \mu \, d\omega,$$

où $d\omega$ est l'angle orienté sous lequel, du point $(x, y) \in A$, on voit l'arc *do* de Γ. Cette fonction est harmonique sur A et, en un

point $(x, y) \in \Gamma$, d'abscisse curvilignes, elle a pour limite :

$$\pi\mu(s) + \int_0^{L(\Gamma)} \mu(\sigma) \frac{\alpha(x-\xi) + \beta(y-\eta)}{(x-\xi)^2 + (y-\eta)^2} d\sigma ;$$

en égalant cette limite à $f(x, y)$, on obtient une équation intégrale non homogène où les variables sont s et σ, mais sans paramètre λ.

Henri Poincaré pressentit, dès 1896, le rôle que ce paramètre jouerait dans les résultats ; cette intuition fut confirmée en 1903 par les remarquables travaux du Suédois Ivar Fredholm résumés ci-dessous (cf. chap. 3).

2. Méthode des approximations successives

Supposons A compact, le noyau K continu sur A^2 et, de même, f dans l'espace de Banach C(A) formé des fonctions y continues sur A à valeurs complexes, avec la norme :

$$\|y\| = \sup_A |y|.$$

Au noyau K est associé *l'opérateur intégral* :

$$\mathcal{K} \in \mathcal{L}[\mathcal{C}(A), \mathcal{C}(A)]$$

qui à la fonction $y \in C(A)$ fait correspondre $z = \mathcal{K}y \in C(A)$ définie par :

$$\text{(3)} \qquad z(x) = \int_A K(x, \xi) y(\xi)\, d\xi.$$

1 désignant l'application identique, on peut écrire l'équation intégrale (1) :

$$\text{(1')} \qquad (I - \lambda\mathcal{K})y = f,$$

et sa résolution revient à inverser l'opérateur $I - \lambda\mathcal{K}$; or (3) permet de déterminer la norme de l'opérateur $\mathcal{K}$:

$$\|\mathcal{K}\| = \sup_{x \in A} \int_A |K(x, \xi)|\, d\xi.$$

Supposons maintenant $\lambda \|\mathcal{K}\| < 1$. D'une part, l'équation homogène :

$$\text{(2')} \qquad y = \lambda\mathcal{K}y$$

ne peut avoir que la solution identiquement nulle ; d'autre part, la suite y_n définie, à partir de $y_0 \equiv 0$ par exemple, par la formule de récurrence ou d'approximations successives (cf. espaces MÉTRIQUES) :

$$y_n = f + \lambda\mathcal{K}y_{n-1},$$

converge dans $\mathcal{C}(A)$ vers la solution de (1), à savoir :

$$\text{(4)} \qquad y = f + \sum_{n \in \mathbf{N}} \lambda^{n+1}\mathcal{K}^{n+1} f,$$

$\mathcal{K}^{n+1}$ étant le $(n + 1)$-ième itéré de l'opérateur $\mathcal{K}$, ou l'opérateur intégral associé au noyau itéré $K^{(n+1)}$ défini par récurrence par :

$$K^{(1)} = K$$

et, pour $n \geqslant 1$, par :

$$\text{(5)} \quad K^{(n+1)}(x, \xi) = \int_{A^n} K(x, t_1) K(t_1, t_2) \ldots \times K(t_{n-1}, t_n) K(t_n, \xi)\, dt_1 \ldots dt_n.$$

On met la solution (4) sous la forme :

$$\text{(6)} \qquad y = (I + \lambda\mathcal{G}_\lambda) f$$

en introduisant l'opérateur intégral $\mathcal{G}_\lambda$ associé au *noyau résolvant :*

$$\Gamma_\lambda = \sum_{n \in \mathbf{N}} \lambda^n K^{(n+1)} ;$$

puisque (6) donne la solution unique de (1), les opérateurs $I - \lambda\mathcal{K}$ et $I + \lambda\mathcal{G}_\lambda$ sont inverses l'un de l'autre, d'où résulte, pour $\lambda \|\mathcal{K}\| < 1$ et $|\mu| \|\mathcal{K}\| < 1$, la relation fondamentale entre noyaux résolvants :

$$\text{(7)} \quad (\mathcal{G}_\lambda - \mathcal{G}_\mu)/(\lambda - \mu) = \mathcal{G}_\lambda \mathcal{G}_\mu = \mathcal{G}_\mu \mathcal{G}_\lambda, \qquad \lambda \neq \mu.$$

Vito Volterra étudia le cas particulier A = [a, b], $a < b$, $K(x, \xi) = 0$ pour $\xi > x$, $K(x, \xi)$ fonction continue de (x, ξ) pour $a \leqslant \xi \leqslant x \leqslant b$. Dans ce cas, l'équation (1) s'écrit :

$$(8) \quad y(x) - \lambda \int_a^x K(x, \xi) y(\xi) \, d\xi = f(x),$$

et la définition (5) des noyaux itérés entraîne :

$$|K^{(n+1)}(x, \xi)| \leqslant \frac{(x-\xi)^n}{n!} (\sup_{A^2} |K|)^{n+1} \quad \text{pour } \xi \leqslant x,$$

$$K^{(n+1)}(x, \xi) = 0 \quad \text{pour } \xi > x.$$

C'est donc pour tout $\lambda \in$ C, et non plus seulement pour $\lambda \, \| \mathcal{K} \| < 1$ comme dans le cas général, que, d'une part, l'équation homogène n'a que la solution identiquement nulle, puisqu'elle entraîne $y = \lambda^n \mathcal{K}^n y$ pour tout n ; que, d'autre part, la série (4) converge dans C(A) vers la solution de (8) ; enfin. on a la relation (7) quels que soient λ et μ.

Les difficultés que présente le cas général furent surmontées par les deux mémoires de Fredholm de 1901 et de 1903.

3. Méthode de Fredholm

Supposons toujours A compact et le noyau K continu sur A'. Si l'on partage A en p parties A, de mesures α_i, $i = 1, \ldots, p$, et si l'on choisit $x_i \in A_i$ pour chaque indice i, on peut considérer le système linéaire :

$$y(x_i) - \lambda \sum_{j=1}^{p} \alpha_j K(x_i, x_j) y(x_j) = f(x_i), \quad i = 1, \ldots, p,$$

comme une approximation de (1) ; or son déterminant est un polynôme en A de degré $\leqslant p$, dont le terme de degré n, $1 \leqslant n \leqslant p$, a pour limite, quand le plus grand des diamètres des α_i, tend vers 0, le monôme :

$$(9) \quad \frac{(-\lambda)^n}{n!} \int_{A^n} K\begin{pmatrix} x_1 & \ldots & x_n \\ x_1 & \ldots & x_n \end{pmatrix} dx_1 \ldots dx_n, \quad n \geqslant 1,$$

où la notation de Fredholm :

$$K\begin{pmatrix} x_1 & \ldots & x_n \\ \xi_1 & \ldots & \xi_n \end{pmatrix}$$

désigne le déterminant des n^2 fonctions $K(x_i, \xi_j)$, i et $j = 1, \ldots, n$.

Le monôme (9) est le terme général d'une série entière convergente pour tout A $\in$ C ; en ajoutant à la série un terme constant égal à 1, on obtient une fonction entière D(A) appelée *déterminante* du noyau K et aussi déterminante du noyau transposé $\tilde{K}$ défini par :

$$(10) \qquad \tilde{K}(x, \xi) = K(\xi, x).$$

Le produit de cette déterminante et du noyau résolvant est encore une fonction entière de A :

$$D(x, \xi; \lambda) = D(\lambda)\Gamma_\lambda(x, \xi) = \sum_{n \geqslant 1} \frac{(-\lambda)^n}{n!} \int_{A^n} K\begin{pmatrix} x x_1 & \ldots & x_n \\ \xi x_1 & \ldots & x_n \end{pmatrix} dx_1 \ldots dx_n + K(x, \xi).$$

Le premier théorème de Fredholm affirme que, si D(h) $\neq 0$, pour tout second membre $f \in$ C(A), l'équation (1) a une solution unique, encore donnée par (6), avec maintenant :

$$\Gamma_\lambda(x, \xi) = \frac{D(x, \xi; \lambda)}{D(\lambda)};$$

c'est donc une fonction méromorphe du paramètre λ, et l'on a la relation (7) quels que soient A et μ.

Si au contraire λ est *valeur singulière* du noyau K, c'est-à-dire D(A) = 0, le deu-

xième théorème de Fredholm affirme que chacune des équations homogènes (2) et :

$$(\tilde{2}) \qquad z(x) = \lambda\int_A K(\xi, x) z(\xi)\, d\xi$$

a des solutions formant un espace vectoriel de dimension finie $d(\lambda) > 0$ commune aux deux équations et au plus égale à l'ordre de multiplicité de la racine A de la déterminante.

Si y est solution de (1) et z solution de (2), on a :

$$\int_A f(x) z(x)\, dx = \int_A y(x) z(x)\, dx - \lambda \int_{A^2} K(\xi, x) y(x) z(\xi)\, dx\, d\xi = 0,$$

d'où $d(\lambda)$ conditions linéaires que le second membref'de (1) doit vérifier pour que (1) ait une solution ; le troisième théorème de Fredholm affirme que ces $d(\lambda)$ conditions nécessaires sont aussi suffisantes.

De ces trois théorèmes se dégage *l'alternative de Fredholm* :

1. Ou bien l'opérateur $I - \lambda\mathcal{K}$ est inversible dans $\mathcal{L}[\mathcal{C}(A), C(A)]$;

2. Ou bien l'opérateur I -AS. n'est ni injectif ni surjectif, son noyau étant de dimension finie, son image étant fermée et de codimension finie.

Le deuxième cas se présente pour les valeurs de λ qui annulent la déterminante : si donc il y en a, elles sont en nombre fini ou forment une suite $\lambda_m \rightarrow \infty$, chaque éventualité pouvant se présenter.

4. Principaux cas particuliers

Dans le *cas de Volterra* (cf. *supra,* fin du chap. 2), la déterminante **ne** peut s'annuler. On retrouve ce fait en remarquant que :

$$K\begin{pmatrix} x_1 & \dots & x_n \\ x_1 & \dots & x_n \end{pmatrix} = K(x_1, x_1)\dots K(x_n, x_n),$$

si $x_1, \dots, x_n$ sont deux à deux distincts. d'où :

$$D(\lambda) = \exp\left[-\lambda \int_a^b K(x, x)\, dx\right]$$

Le *cas de Goursat* est celui où le noyau K est de la forme :

$$K(x, \xi) = \sum_{i=1}^{p} h_i(x) k_i(\xi),$$

les h_i étant linéairement indépendantes dans C(A), ainsi que les k_i. Dans ce cas, on a

$$K\begin{pmatrix} x_1 & \dots & x_n \\ x_1 & \dots & x_n \end{pmatrix} \equiv 0,$$

pour $n > p$; dont D(A) est un polynôme de degré $\leqslant p$ et le noyau K a au plus p valeurs singulières.

Le noyau K est dit *hermitien* si :

$$\tilde{K} \equiv \overline{K},$$

c'est-à-dire si :

$$K(\xi, x) = \overline{K}(x, \xi),$$

quels que soient x et $\xi \in A$. Dans ce cas, chaque déterminant de Fredholm :

$$K\begin{pmatrix} x_1 & \dots & x_n \\ x_1 & \dots & x_n \end{pmatrix}$$

est réel et les valeurs singulières aussi ; car :

$$y = \lambda \mathcal{K} y$$

entraîne la relation :

$$\int_A |y(x)|^2\, dx = \lambda \int_{A^2} K(x, \xi) \bar{y}(x) y(\xi)\, dx\, d\xi,$$

où les deux intégrales sont réelles. Le noyau K a certainement au moins une valeur singulière s'il n'est pas identiquement nul, et il a certainement une suite λ_m, $m \in \mathbb{N}$, de telles valeurs singulières, si le noyau n'est pas un noyau de Goursat. Le

cas particulier de Goursat étant exclu, on a les formules :

$$(11a)\quad \sum_{m\in N} \frac{1}{\lambda_m^n} = \int_A K^{(n)}(x,x)\,dx,\ n \geqslant 2,$$

et :

$$(11b)\quad D(\lambda) = \exp\left[-\lambda \int_A K(x,x)\,dx\right] \times \prod_{m\in N}\left(1-\frac{\lambda}{\lambda_m}\right)\exp\frac{\lambda}{\lambda_m}\cdot$$

La richesse et la précision des résultats obtenus dans cette théorie, pour des noyaux quelconques et plus encore pour des noyaux hermitiens, ne sont pas seulement admirables par elles-mêmes : elles ont contribué largement à l'essor, au XXe siècle, de l'analyse fonctionnelle en général, à l'essor de la théorie des opérateurs linéaires et à celui de la théorie des espaces préhilbertiens ou hilbertiens en particulier.

5. Opérateurs compacts

Propriété de compacité

L'inégalité de Schwarz, appliquée à (3), donne :

$$|z(x)|^2 \leqslant \int_A |y(\xi)|^2 d\xi \int_A |K(x,\xi)|^2 d\xi.$$

et :

$$|z(x')-z(x)|^2 \leqslant \int_A |y(\xi)|^2 d\xi \int_A |K(x,\xi)-K(x',\xi)|^2 d\xi.$$

Ces inégalités suggèrent les hypothèses suivantes sur le noyau K : l'espace C'(A) contient chaque fonction :

$$K_x : \xi \mapsto K(x,\xi)$$

et l'application $x \rightarrow K_x$ est une application continue de A dans C'(A).

Sous ces hypothèses, réalisées en particulier si K est continu sur le compact A', que l'on munisse C(A) de la norme :

$$\|y\| = \sup_A |y|$$

comme aux chapitres 2 et 3, ou que l'on munisse C(A) de la norme :

$$\|y\| = \left[\int_A |y(\xi)|^2 d\xi\right]^{1/2}$$

définissant une topologie strictement moins fine, l'opérateur intégral $\mathcal{K}$ a cette propriété (obtenue en appliquant le théorème d'Ascoli à la suite bornée équicontinue $\mathcal{K}y_n$) que, pour toute suite y,, bornée dans C(A), la suite $\mathcal{K}y_n$ contient une suite partielle convergente dans C(A).

Cette propriété de l'opérateur $\mathcal{K}$ fut dégagée, puis étudiée dans un espace vectoriel normé quelconque E, par le Hongrois Frédéric Riesz, sous le nom de *complète continuité.* auquel on préfère aujourd'hui celui de *compacité :* elle entraîne en effet la continuité de l'opérateur, mais s'oppose à ce qu'il ait un inverse continu, du moins si E est de dimension infinie.

La seconde norme sur C(A) indiquée ci-dessus a sur la première l'avantage de munir C(A) d'une structure préhilbertienne (cf. espace de HILBERT, chap. 1) permettant de considérer d'autre part des opérateurs autoadjoints (cf. ci-dessous).

Valeurs spectrales

Soit E un espace vectoriel normé quelconque et $\mathcal{K} \in \mathcal{L}(E, E)$: on dit qu'un nombre complexe ζ est *valeur spectrale* de $\mathcal{K}$ si l'opérateur $\mathcal{K} \quad \zeta I$ n'est pas inversible dans C(E, E), *valeurpropre* de $\mathcal{K}$ si $\mathcal{K} \quad \zeta I$ n'est pas injectif ; ceci entraîne cela, et réciproquement, si E est de dimension finie.

Soit E de dimension infinie et $\mathcal{K}$ compact : alors, la valeur 0 est spectrale, mais elle n'est pas propre en général ; au contraire, si $\zeta \neq 0$, on a pour l'opérateur $\mathcal{K} - \zeta\mathbf{I}$ l'alternative de Fredholm telle qu'elle a été énoncée à la fin du chapitre 3, de sorte qu'une valeur $\zeta \neq 0$ est propre si et seulement si elle est spectrale. S'il y a de telles valeurs. elles sont en nombre fini ou forment une suite tendant vers 0.

Opérateur adjoint

Supposons désormais que la norme de E lui donne une structure préhilbertienne, donc qu'elle est associée à un produit scalaire, noté $(x \mid y)$; on dit alors que deux opérateurs $\mathcal{K}$ et $\mathcal{K}^* \in$ C(E, E) sont *adjoints* si :

$$(\mathcal{K}x \mid y) = (x \mid \mathcal{K}^* y),$$

quels que soient x et $y \in$ E, et que $\mathcal{K}$ est *autoadjoint* s'il est son propre adjoint. Ainsi, lorsque E est égal à $\mathcal{C}(A)$ muni de la seconde norme indiquée *supra* (cf. *Propriété de compacité*), les opérateurs intégraux associés aux noyaux K et $K^* = \tilde{\bar{K}}$ sont adjoints.

Si $\mathcal{K}$ et $\mathcal{K}^*$ sont adjoints et compacts, on retrouve le troisième théorème de Fredholm sous la forme suivante : L'image de l'opérateur $\mathcal{K} - \zeta\mathbf{I}$ est le supplémentaire orthogonal du noyau de son adjoint $\mathcal{K}^* - \bar{\zeta}\mathbf{I}$.

Soit enfin $\mathcal{K}$ compact autoadjoint : ses valeurs spectrales sont réelles, les noyaux de $\mathcal{K} - \zeta\mathbf{I}$ et $\mathcal{K} - \zeta'\mathbf{I}$ sont orthogonaux si $\zeta \neq \zeta'$ et si $\|\mathcal{K}\|$ ou $-\|\mathcal{K}\|$ est valeur spectrale.

Les valeurs spectrales peuvent former un ensemble fini $\{\zeta_0, \zeta_1, \ldots, \zeta_n\}$, avec $\zeta_0 = 0$. Alors, le supplémentaire orthogonal du sous-espace engendré par les noyaux des :

$$\mathcal{K} - \zeta_m\mathbf{I}, \quad m = 1, \ldots, n$$

est stable par $\mathcal{K}$ et la restriction de $\mathcal{K}$ à ce supplémentaire est un opérateur compact autoadjoint sans autre valeur spectrale que 0, donc nul ; cela veut dire que tout vecteur $y \in$ E est la somme de ses projections orthogonales y_m sur les noyaux des $\mathcal{K} - \zeta_m\mathbf{I}$, $m = 0, 1, \ldots, n$, d'où les formules :

$$(12) \qquad y = \sum_{m=0}^{n} y_m, \quad \mathcal{K}y = \sum_{m=1}^{n} \zeta_m y_m,$$

qui définissent parfaitement l'opérateur $\mathcal{K}$ et permettent, par exemple, de traiter l'équation :

$$(\mathcal{K} - \lambda\mathbf{I})y = z,$$

où z est donné : si λ n'est pas valeur spectrale, l'équation a pour solution unique :

$$(13) \qquad y = \sum_{m=0}^{n} \frac{z_m}{\zeta_m - \lambda},$$

où z_m est la projection orthogonale de z sur le noyau de :

$$\mathcal{K} - \zeta_m\mathbf{I};$$

si, au contraire, λ est une valeur spectrale ζ_μ, l'équation n'a de solutions que si $z_\mu = 0$; ces solutions sont encore données par (13), à ceci près que le terme de rang μ est remplacé par un élément quelconque du noyau de :

$$\mathcal{K} - \zeta_\mu\mathbf{I}.$$

Les valeurs spectrales de $\mathcal{K}$ peuvent aussi former une suite ζ_m, tendant vers $\zeta_0 = 0$, et les formules (12) et (13) subsistent à condition que les séries qu'elles contiennent convergent dans E : mis à part le cas $\lambda = 0$, il suffit pour cela que E soit complet ou soit un espace de Hilbert. On remarque, à ce sujet, que C(A) n'est pas complet : pour un opérateur intégral $\mathcal{K}$

dans C(A), associé à un noyau K hermitien mais non noyau de Goursat, donc possédant une suite de valeurs spectrales, on a pourtant le remarquable théorème de Hilbert-Schmidt, d'après lequel la deuxième série (12) converge uniformément vers la fonction $\mathcal{K}y$.

MICHEL HERVÉ

Bibliographie

C. CORDUNEANU, *Integral Equations and Applications*, Cambridge Univ. Press. New York, 1991 / R. COURANT & D. HILBERT, *Methods of mathematical Physics*, vol. 1. J. Wiley, New York, 1989 / J. DIEUDONNÉ. *Éléments d'analyse*. 9 vol.. Gauthier-Villars. Paris, 1975-1983 / E. GOURSAT, *Cours d'analyse mathématique*, vol. III, fac-sim. de l'éd. Gauthier-Villars, 1923-1925, Gabay Sceaux, 1992 / J. GUY & A. SALÈS, *Integral Equations in Everyday Practice*, Lavoisier, Paris. 1991 / H. HOCHSTADT, *Integral Equations*, Wiley, New York, 1989 / M. KRASNOV et al., *Équations intégrales*, M.I.R., Moscou-Paris, 1977 / R. KRESS, *Linear Integral Equations*, Springer-Verlag, New York, 1989 / D. PORTER & D. S. STIRLING, *Integral Equations A Practical Treatment, from Spectral Theory to Applications*, Cambridge Univ. Press, 1991 / F. RIESZ & B. Sz. NAGY, *Leçons d'analyse fonctionnelle*, fac-sim. de l'éd. Gauthier-Villars, Paris, 1955, Gabay, 1990 / A. F. RUSTON, *Fredholm Theory in Banach Spaces*, Cambridge Univ. Press, New York, 1986 / V. VOLTERRA, *Leçons sur les équations intégrales et intégro-différentielles*, *ibid.*, Paris. 1913.

INTÉGRATION & MESURE

La théorie de l'intégration joue en mathématique un rôle extrêmement important. C'est une théorie riche et complexe. Il ne sera pas question ici d'en donner une description exhaustive ni d'en aborder les assez redoutables aspects techniques. On s'efforcera de mettre en lumière les grandes idées simples qui y sont à l'œuvre et de montrer comment elles lient les aspects les plus élémentaires de la théorie de la mesure aux développements les plus généraux de la théorie de l'intégration.



1. Le problème initial

Généralités

Mesurer est une activité dont l'existence est attestée dans toutes les sociétés historiques, et il est assez surprenant de constater que ce n'est que dans un passé relativement récent, au début du XX[e] siècle, que la réflexion mathématique a commencé à en établir une théorie claire et cohérente.

Il faut tout de suite remarquer que le problème, pris dans sa plus grande généralité, reste encore mystérieux : on ne sait pas encore très clairement, par exemple, le sens qu'il faut donner au mot « mesurer », ni même si on peut vraiment lui donner un sens dans le domaine des sciences humaines, bien qu'on ait l'impression qu'une bonne conception de la mesure puisse y être l'origine de progrès décisifs.

Il faut encore remarquer que le mot « mesure » a, en mathématique, un sens plus restreint qu'en physique, et que ce que le mathématicien appelle théorie de la mesure ne s'applique directement qu'à une partie de l'activité du physicien et ne vise que la structure conceptuelle à l'exclusion des procédés expérimentaux de détermination des valeurs numériques qui relèvent de la métrologie.

Ce qui a retardé la naissance d'une bonne théorie de la mesure a été l'incapacité où est demeurée longtemps l'humanité de distinguer nettement entre ce que

l'on mesure d'une part et l'échelle avec laquelle on mesure d'autre part, et de concevoir clairement ce qui les lie. L'échelle est constituée par le corps ordonné R des nombres réels, dont la théorie définitive n'a été élaborée qu'à la fin du XIXe siècle (G. Cantor, R. Dedekind), mais qui avait été déjà presque totalement construit par le mathématicien grec Eudoxe, au IVe siècle avant J.-C., sous le nom, que la tradition a conservé, de mesure des grandeurs (et de rapports de grandeurs), qui est révélateur de la confusion signalée plus haut. En réalité, Eudoxe ne considérait que ce que nous appelons les nombres réels positifs. Il le faisait d'une manière rigoureusement correcte et extrêmement pure : il est émouvant pour le mathématicien contemporain de constater le souci, très moderne, qu'a montré Eudoxe de ne rien introduire qu'il ne puisse construire par combinaison de concepts préalablement bien délimités. Mais sa construction, impeccable au point de vue de la rigueur, en ce qui concerne l'édification d'une échelle de valeurs, était beaucoup moins claire en ce qui concerne les « grandeurs ». Elle était en outre d'une complexité qui l'a rendue inutilisable pour la plupart de ses successeurs ; ceux-ci ne l'ont pas comprise et n'en ont retenu que des bribes, qu'ils se sont transmises avec une telle persévérance qu'on les trouve encore à l'état de vestiges peu intelligibles dans la plupart des manuels élémentaires. Ce qui a manqué à Eudoxe, c'est une autre idée très moderne qui consiste, lorsque l'on a construit, à partir d'un matériel conceptuel initial, de nouveaux êtres mathématiques très complexes par rapport aux éléments initiaux, à considérer les propriétés essentielles de ces nouveaux êtres et à repartir en les prenant à leur tour comme éléments initiaux d'une nouvelle construction. Dans la construction d'Eudoxe, comme dans celle de Dedekind, un réel apparaît comme un ensemble de rationnels, chaque rationnel étant lui-même un ensemble de couples d'entiers ; mais Eudoxe reste fidèle au langage des entiers, tandis que Dedekind dégage explicitement la structure du corps totalement ordonné archimédien et complet des réels. (À noter d'ailleurs que l'axiome dit d'Archimède est explicitement attribué à Eudoxe par Archimède lui-même.)

C'est l'apparition presque simultanée d'une bonne théorie des nombres réels et de la notion d'ensemble qui a créé les conditions favorables à la naissance de la théorie moderne de la mesure.

Formulation de la question

Reprenons la question dans un cas simple : tout le monde a appris à calculer la surface ou l'aire de certaines régions du plan, et les mathématiciens des siècles passés ont consacré beaucoup d'efforts à calculer les aires de régions de plus en plus compliquées, sans jamais cependant dire très explicitement pourquoi ils menaient leurs calculs comme ils le faisaient, ni ce qu'ils attendaient du résultat. Expliciter les idées non formulées qui présidaient à ces recherches n'eut pas pour seul effet de satisfaire les exigences de rigueur et d'esthétique du mathématicien, cela lui permit de forger les outils propres à déterminer, puis à étendre le domaine de validité de ces calculs et de les effectuer dans tous les cas où ils sont valables.

Qu'attendons-nous en effet de ces calculs d'aire ? Ils permettent (une unité d'aire étant choisie) d'attribuer à chaque région du plan d'un certain type un nombre réel positif que l'on appelle son aire, la propriété essentielle étant souvent exprimée par le fait que l'on peut « ajouter » des

aires, ce que l'on peut décrire sous une première forme en disant que, si une région R apparaît comme formée de deux régions R_1 et R_2 « qui n'empiètent pas l'une sur l'autre », l'aire de R est la somme des aires de R_1 et de R_2. Deux régions telles que l'on puisse les « transporter » (au moins idéalement) l'une sur l'autre ont des aires égales.

Si l'on veut préciser ces idées, il faut d'abord savoir ce que l'on entend par « région » du plan. Le plan étant considéré comme un ensemble dont les éléments sont les points, le projet le plus ambitieux sera de considérer que tout sous-ensemble du plan doit avoir une aire. Cela signifierait donc que l'on peut définir une application *m* (appelée mesure universelle) de l'ensemble Y(P) de toutes les parties du plan P dans l'ensemble des nombres réels positifs, et que cette application a les propriétés d'additivité et d'invariance par isométrie.

Additivité. Pour tout couple (A, B) de parties disjointes du plan (c'est-à-dire telles que $A \cap B = \emptyset$), on a :

$$m(A \cup B) = m(A) + m(B).$$

Cette condition peut être également formulée sous la forme équivalente suivante :

$$m(\emptyset) = 0,$$

et alors, pour tout couple de parties (A, B), du plan :

$$m(A \cup B) + m(A \cap B) = m(A) + m(B);$$

ce qui permet de préciser le sens de « ne pas empiéter l'un sur l'autre » par « avoir une intersection d'aire nulle ».

Invariance par isométrie. Quelle que soit la partie A du plan, et quelle que soit l'isométrie τ du plan (une isométrie est une bijection du plan sur le plan qui conserve la distance), on a :

$$m[\tau(A)] = m(A).$$

Cette définition précise étant donnée, deux problèmes se posent : Existe-t-il de telles applications ? Si oui, en existe-t-il une seule, ou plusieurs ? Il est clair que, s'il en existe une, que nous appellerons *m*, et, si λ est un réel positif fixe, λm en est une autre. L'unicité recherchée ne sera donc qu'une unicité à un facteur près, Cela dit, la réponse aux questions posées est si peu évidente qu'on a, en réalité, les résultats suivants. Il existe une telle application dans le cas du plan, mais l'unicité n'est pas assurée. Il n'existe pas de telle application dans l'espace euclidien à trois dimensions (où l'on parle alors de volume et non plus d'aire), ni pour les espaces euclidiens à plus de trois dimensions.

Espaces mesurés

La non-existence ou la non-unicité amènent à restreindre nos ambitions initiales et à reposer le problème en essayant de définir l'application non pas sur l'ensemble de toutes les parties du plan ou de l'espace, mais sur un sous-ensemble. Pour donner une formulation plus générale, nous allons abandonner le plan et supposer que l'on part d'un ensemble X quelconque : Sur quelles parties $\mathcal{A}$ de T(X) est-il raisonnable de chercher à définir une application additive ?

Une première exigence est que, si A et B sont deux parties disjointes de X qui appartiennent à $\mathcal{A}$, leur réunion appartient aussi à $\mathcal{A}$. Mais il est plus commode d'exiger que, quelles que soient les parties A et B éléments de $\mathcal{A}$, alors $A \cup B$ soit aussi élément de $\mathcal{A}$. L'ensemble $A \cup B$ est d'autre part la réunion de A et de B − A,

qui sont disjoints, et il est commode que B — A appartienne aussi à $\mathcal{A}$.

Ce qui s'est effectivement révélé être la bonne structure pour l'ensemble $\mathcal{A}$ est exprimé par la condition suivante : Pour tout couple (A, B) d'éléments de $\mathcal{A}$, alors A U B et A B sont éléments de $\mathcal{A}$. (On peut montrer que cette condition est équivalente au fait que la différence symétrique A △ B et l'intersection A ∩ B sont éléments de $\mathcal{A}$, si A et B le sont.) On dit alors que $\mathcal{A}$ est un *clan* de parties de X, ou quelquefois un anneau de Boole ($\mathcal{A}$ a une structure d'anneau pour les opérations △ et ∩ (cf. ANNEAUX ET ALGÈBRES, chap. 2).

On dira alors qu'une application additive m de $\mathcal{A}$ dans l'ensemble $\mathbf{R}^+$ des nombres réels positifs est une *mesure* (en précisant, si besoin est, simplement additive) ou encore une étendue (*content*, *Inhalt*). La situation fondamentale est le triplet (X, il, m), où $\mathcal{A}$ est un clan de parties de X, et m une application additive de $\mathcal{A}$ dans $\mathbf{R}^+$. Un tel triplet est souvent appelé un *espace mesuré*.

Dans certains cas, il peut être intéressant (comme dans le cas de l'aire et du volume) d'avoir une propriété d'invariance par rapport à un sous-groupe G du groupe des bijections de X sur X, ce qui signifie que, pour tout g ∈ G et tout A ∈ $\mathcal{A}$, on ag(A) ∈ $\mathcal{A}$ et $m[g(A)] = m(A)$, mais la théorie ne se limite pas à ce cas.

L'exemple le plus ancien d'espace mesuré qui ait été considéré par l'humanité est celui où X est un ensemble, $\mathcal{A}$ l'ensemble de ses parties finies, et $m(A)$ le nombre d'éléments de la partie finie A, et l'on peut dire que c'est à partir de ce cas, par extensions successives, qu'a été développée toute la théorie de la mesure.

Notons que, l'intersection d'une famille de clans étant un clan, il existe toujours un plus petit clan contenant une partie donnée de T(X) : on dit alors que ce clan est engendré par la partie. On peut montrer qu'il existe des mesures non triviales (c'est-à-dire ne prenant pas uniquement la valeur 0) sur tout clan.

Voici des exemples usuels et importants d'espaces mesurés.

(a) X = R, et A est l'ensemble des réunions finies de semi-segments $[a, b[$ ($a \leqslant x < b$), avec $m([a, b[) = b \quad a$. Il y a invariance par rapport aux translations de R. Cette mesure n'est autre que la longueur. On notera qu'il est préférable de considérer la longueur comme une application de $\mathcal{A}$ dans $\mathbf{R}^+$, plutôt que de parler de la mesure de la longueur, qui, ou bien est un pléonasme, ou bien se réfère à un concept de longueur défini indépendamment, ce qui complique inutilement la situation.

(b) X est un segment $[\alpha, \beta]$ de la droite, ct $\mathcal{A}$ est le clan engendré par les semi-segments [a, b[et les segments $[a, \beta]$. On définit m au moyen d'une fonction positive croissante g définie sur $[\alpha, \beta]$ en posant $m([a, x[) = g(x)$ et $m([\alpha, \beta]) = g(\beta)$. On peut interpréter m comme une masse. Si g n'est pas affine, cette masse m n'est pas « uniformément répartie ».

(c) X est le plan et $\mathcal{A}$ est l'ensemble des réunions finies de rectangles dont les côtés ont des directions fixes, et la mesure d'un rectangle est le produit des longueurs de ses côtés. Il y a invariance par rapport au groupe des translations du plan.

(cl) X est un ensemble d'éventualités, $\mathcal{A}$ un clan d'événements, avec X ∈ A, et m une probabilité, à qui est imposée la condition m(X) = 1.

Les exemples les plus classiques représentés par ce schéma concernent, comme ci-dessus, les longueurs, les aires, les volumes, les masses et les probabilités : l'inclusion de la théorie des probabilités dans la théorie de la mesure (A. Kolmogoroff) a

été l'origine du développement moderne de la première et d'un considérable enrichissement de la seconde.

Extension d'une mesure

La formulation moderne du problème que les mathématiciens s'efforçaient de résoudre en calculant les mesures d'ensembles de plus en plus complexes est la suivante : Étant donné un espace mesuré (X, $\mathcal{A}$, m), est-il possible de déterminer d'autres triplets (X, $\hat{\mathcal{A}}$, $\hat{m}$), où $\mathcal{A} \subset \hat{\mathcal{A}}$ et où $\hat{m}$ est une extension de m à $\hat{\mathcal{A}}$?

L'idée d'une solution remonte-t-elle aussi à Eudoxe ? Soit A une partie de X n'appartenant pas à $\mathcal{A}$, mais telle qu'il existe au moins un élément $\beta \in \mathcal{A}$ avec $A \subset \beta$. On considère alors tous les éléments $a \in \mathcal{A}$ tels que $\alpha \subset A$ (il en existe : il y a au moins l'ensemble vide) et tous les éléments $\beta \in \mathcal{A}$ tels que $A \subset \beta$. On a donc :

$$\alpha \subset A \subset \beta.$$

Or, m étant additive et positive, m est croissante (c'est-à-dire que $a \subset \beta$ entraîne $m(a) \leqslant m(\beta)$). On en déduit alors que la borne supérieure de l'ensemble des nombres $m(a)$ est au plus égale à la borne inférieure de l'ensemble des nombres $m(\beta)$. L'idée est que l'on pourra étendre m aux ensembles pour lesquels ces deux bornes sont égales, et le résultat qui justifie tous les calculs classiques est le suivant : La famille des ensembles A pour lesquels la borne supérieure des $m(a)$ et la borne inférieure des $m(\beta)$ sont égales est un clan $\hat{\mathcal{A}}$ qui contient $\mathcal{A}$, et, en posant :

$$\hat{m}(A) = \sup m(a),$$

on obtient une mesure sur $\hat{\mathcal{A}}$ dont la restriction à $\mathcal{A}$ est m. En général, $\hat{\mathcal{A}}$ est strictement inclus dans $\mathcal{P}(X)$, et l'application du même procédé à (X, $\hat{\mathcal{A}}$, $\hat{m}$) redonnerait le même triplet.

Dans le cas des aires planes, partant de l'espace mesuré de l'exemple (*c*), on trouve tous les ensembles qualifiés classiquement de *quarrables*, et l'espace mesuré est invariant pour les isométries du plan : on obtient l'aire classique la plus générale.

2. Linéarisation et intégrale de Riemann

Soit (X, $\mathcal{A}$, m) un espace mesuré. À chaque élément A de $\mathcal{A}$, associons sa fonction caractéristique φ_A et considérons les combinaisons linéaires à coefficients réels de ces fonctions caractéristiques : on obtient des fonctions dites étagées (relativement à $\mathcal{A}$) et leur ensemble V a une structure naturelle d'espace vectoriel réticulé. Si φ et ψ sont deux éléments de V, $\sup(\varphi, \psi)$ et $\inf(\varphi, \psi)$ appartiennent aussi à V.

On peut alors associer à m une forme linéaire 1 sur V, en posant :

$$I(\phi) = \sum \lambda_i m(A_i)$$

pour :

$$\varphi = \sum \lambda_i \varphi_{A_i},$$

et l'on vérifie que, si φ s'exprime de deux manières différentes comme combinaisons linéraires de fonctions φ_{A_i}, on obtient bien, dans les deux cas, la même valeur pour $I(\varphi)$. La linéarité de I est évidente. En outre, si $\varphi \geqslant 0$, on a $I(\varphi) \geqslant 0$.

On peut alors poser un problème d'extension de la forme I à un espace vectoriel contenant V. Le procédé classique de l'intégration de Riemann est l'analogue du procédé d'Eudoxe pour l'extension des mesures et peut être ainsi décrit. On considère les fonctionsfdéfinies sur X à valeurs dans R, qui ont la propriété

d'être bornées et de s'annuler hors d'un ensemble $A_f \in \mathcal{A}$. On peut encadrerfpar des fonctions étagées φ et ψ telles que $\varphi < f < \psi$. On a donc :

$$I(\varphi) \leqslant I(\psi),$$
$$\sup\{I(\varphi) \mid \varphi \in V, \varphi < f\} \leqslant \inf\{I(\psi) \mid \psi \in V, f < \psi\} ;$$

on montre que l'ensemble des fonctions pour lesquelles l'égalité a lieu dans la formule précédente est un espace vectoriel réticulé V qui contient V, et que, si l'on prend pour $\hat{I}(f)$ la valeur commune aux deux bornes, on définit sur $\hat{V}$ une forme linéaire positive $\hat{I}$ qui prolonge I.

Tel est l'essentiel de l'intégration au sens de Riemann. Remarquons que, partant de (X, $\mathcal{A}$, m), la famille des ensembles A dont les fonctions caractéristiques φ appartiennent à Y n'est autre que $\hat{\mathcal{A}}$, et que $\hat{I}(\varphi)$ n'est alors autre que $\hat{m}(A)$. On ne perd donc rien, et on gagne beaucoup à procéder à cette linéarisation et à travailler sur des espaces vectoriels de fonctions et sur leurs formes linéraires positives.

Un problème technique concernant cette intégration est de savoir, suivant l'espace mesuré (X, $\mathcal{A}$, m), ou l'espace V, dont on part, si l'on peut caractériser indépendamment de l'intégration les fonctions de $\hat{V}$: dans cet ordre d'idées, si X est localement compact, et si $\mathcal{A}$ contient tous les compacts de X, toutes les fonctions continues à support compact appartiennent à V. Plus particulièrement, si (X, $\mathcal{A}$, m) est l'espace mesuré de l'exemple (b), où m est définie à partir d'une fonction croissante g, toutes les fonctions continues réelles sur $[\alpha, \beta]$ sont intégrables, et l'intégrale est appelée l'intégrale de Riemann-Stieltjes par rapport à g et est souvent notée :

$$\int_\alpha^\beta f\,dg.$$

Une fonction ν à variation bornée étant la différence de deux fonctions croissantes p et n, on peut définir l'intégrale de toute fonction continue f par rapport à v en posant :

$$\int_\alpha^\beta f\,dv = \int_\alpha^\beta f\,dp - \int_\alpha^\beta f\,dn$$

et on a :

$$\left| \int_\alpha^\beta dv \right| \leqslant V(\alpha, \beta)\|f\|,$$

où $\|f\|$ est la norme uniforme de f, définie par :

$$\|f\| = \sup\{|f(x)|,\ x \in [a, b]\} ;$$

$V(\alpha, \beta)$ est la variation absolue de v entre α et β, soit :

$$V(\alpha, \beta) = \sup \sum_{k=1} |f(x_k) - f(x_{k-1})|,$$

où la borne supérieure est prise par rapport à l'ensemble des subdivisions finies $\alpha = x_0 < x_1 < < x_{n-1} < x_n = \beta$ de l'intervalle $[\alpha, \beta]$.

Autrement dit, sur l'espace de Banach des fonctions continues réelles définies sur $[\alpha, \beta]$, les intégrales de Riemann-Stieltjes sont des formes linéaires continues. En 1909, F. Riesz a prouvé qu'elles étaient les seules.

3. La théorie de Lebesgue

L'odditivité dénombrable

On s'est efforcé, dans ce qui précède, de mettre en lumière les idées implicites essentielles de la théorie classique de la mesure et de l'intégration telle qu'elle s'est développée non sans difficultés des Grecs à Riemann, et qui constitue ce que l'on peut appeler la théorie élémentaire de la mesure.

Mais, historiquement, cette prise de conscience de ce qui intervenait fondamentalement dans la théorie classique s'est produite en même temps, sinon plus tard que l'introduction d'une nouvelle idée extrêmement féconde, due à É. Borel, et qui est celle de I'additivité dénombrable.

Reprenons un triplet (X, $\mathcal{B}, p$), et supposons que le clan $\mathcal{B}$ contient non seulement les réunions finies de ses éléments, mais aussi les réunions dénombrables, c'est-à-dire supposons que la réunion de toute famille dénombrable d'éléments de $\mathcal{B}$ soit un élément de $\mathcal{B}$. On dit alors que $\mathcal{B}$ est une *tribu*, ou encore un o-anneau (de Boole).

Supposons que μ soit non seulement additive, mais vérifie la condition suivante : Pour toute famille dénombrable (A_i), $i \in$ N d'éléments de $\mathcal{B}$ deux à deux disjoints, on a :

$$\mu\left(\bigcup_{i \in N} A_i\right) = \sum_{i \in N} \mu(A_i);$$

on dit alors que μ est *dénombrablement additive* (ou o-additive). C'est cette situation qui a été envisagée par Bore1 et qui est toujours, sauf précision limitative, désignée par le terme *espace mesuré*.

Un premier problème est celui de l'existence de tels triplets. Il est évident qu'il existe des tribus : pour tout ensemble X, l'ensemble $\mathcal{P}(X)$ de ses parties est une tribu. De plus, l'intersection de toute famille de tribus étant une tribu, il existe, pour toute partie $\mathcal{C}$ de :I'(X), une plus petite tribu qui la contient ; on appelle cette dernière la tribu engendrée par $\mathcal{C}$.

L'existence d'une mesure o-additive a posé un problème plus redoutable, qui a été résolu par Lebesgue suivant une voie que l'on peut schématiser, dans un cadre plus général que le sien, de la façon suivante : Partons d'un triplet (X, $\mathcal{A}$, m), où $\mathcal{A}$ est un clan et m une mesure simplement additive, et considérons la famille $D(\mathcal{A})$ des parties A de X telle que chacune de ces parties soit incluse dans la réunion d'une famille dénombrable (α_i), $i \in$ N, d'éléments de $\mathcal{A}$. On peut considérer la somme :

$$\sum_{i \in N} m(\alpha_i),$$

qui est un réel positif ou $+\infty$, et désigner par m*(A) la borne inférieure de l'ensemble de ces sommes obtenues en considérant toutes les familles dénombrables d'éléments de $\mathcal{A}$ dont la réunion contient A. L'idée est que m*(A) doit fournir une approximation par le haut (en langage plus mathématique, un majorant) de l'éventuelle mesure de A. On définit ainsi une application m^* de $D(\mathcal{A})$ dans $\overline{\mathbf{R}}_+$ (union de $\mathbf{R}_+$ et de $\{+\infty\}$). Il est clair qu'une réunion dénombrable d'éléments A_i de $D(\mathcal{A})$ est un élément de $D(\mathcal{A})$ et l'on a :

$$m^*\left(\bigcup_{i \in N} A_i\right) \leqslant \sum_{i \in N} m^*(A_i).$$

Cette propriété, plus faible que la o-additivité, est appelée la σ-sous-additivité, et m^* est qualifiée de *mesure extérieure* associée à (X, A, m).

On considère alors les éléments A de D(it) tels que l'on ait, pour tout élément E de $D(\mathcal{A})$:

$$m^*(E) = m^*(A \cap E) + m^*(\complement A \cap E),$$

et on montre que leur ensemble est une tribu $\mathcal{B}$ qui contient $\mathcal{A}$ et que la restriction μ de m^* à $\mathcal{B}$ est o-additive. Les éléments de $\mathcal{B}$ sont qualifiés d'*ensembles mesurables*, ou ($\mathcal{A}$, m)-mesurables si on veut rappeler leur origine.

Lebesgue a d'abord démontré ce résultat dans le cas où X est un segment de la

droite, $\mathcal{A}$ le clan engendré par les segments, et m la longueur. Le même processus peut être appliqué à la droite entière, et la mesure obtenue, qui prolonge la longueur et qui est invariante par translation, est appelée mesure de Lebesgue de la droite. (Le même procédé réussit pour les aires, les volumes et l'on parle encore de mesure de Lebesgue dans $\mathbf{R}^2$, $\mathbf{R}^3$, $\mathbf{R}^n$.)

Il faut noter que $\mathcal{B}$ est en général strictement incluse dans :I'(X) et on démontre par exemple, dans le cas de la droite, qu'il est impossible d'étendre la mesure de Lebesgue à $\mathcal{P}(X)$ tout entier.

La tribu $\mathcal{B}$ dépend de $\mathcal{A}$ et de m, mais $\mathcal{B}$ contient toujours la plus petite tribu contenant $\mathcal{A}$. En particulier, dans le cas de $\mathbf{R}^n$, la tribu des ensembles mesurables au sens de Lebesgue ($\mathcal{A}$ étant le clan engendré par les produits d'intervalles bornés, appelés pavés, et m le volume n-dimensionnel) contient la tribu engendrée par les pavés, qu'on appelle *tribu borélienne* de $\mathbf{R}^n$, elle contient tous les ouverts et tous les fermés de $\mathbf{R}^n$. (Ses éléments sont appelés *boréliens* de $\mathbf{R}^n$.)

Dans tous les cas, la tribu $\mathcal{B}$ est *complète par rupport à* μ, en ce sens que, si $A \in \mathcal{B}$ **avec** $\mu(A) = 0$, **toute** partie de A appartient aussi à $\mathcal{B}$ et a aussi une mesure nulle. Le triplet (X, $\mathcal{B}$, μ) a été construit à partir du triplet (X, $\mathcal{A}$, m) et on a $\mathcal{A}$ C 3 ; μ est-elle une extension de m ? En général, on a seulement pour $\alpha \in \mathcal{A}$:

$$\mu(\alpha)\ 6\ m(\alpha)\ ;$$

l'égalité est assurée pour tout $\alpha \in \mathcal{A}$, si et seulement si, pour tout élément α de $\mathcal{A}$ qui est réunion d'une famille dénombrable d'éléments deux à deux disjoints α_i, de $\mathcal{A}$, on a :

$$m(a) = \sum_{i \in \mathbf{N}} m(\alpha_i)\ ;$$

c'est ce qui est réalisé pour les longueurs, les aires, les volumes...

L'intégrale de Lebesgue

En même temps qu'il démontrait l'existence de mesures o-additives, Lebesgue définissait l'intégrale qui porte son nom.

Dans le cas simple de fonctions réelles bornées nulles hors d'un élément de $\mathcal{B}$ de mesure finie, le processus indiqué pour l'intégrale de Riemann conduit, à condition de partir des fonctions étagées relatives à (X, $\mathcal{B}$, μ), à l'intégrale de Lebesgue. Les fonctions bornées nulles hors d'un ensemble mesurable de mesure finie qui sont intégrables sont alors exactement celles que Lebesgue a appelées mesurables, c'est-à-dire celles qui donnent pour image réciproque de tout borélien de R un ensemble mesurable (c'est-à-dire appartenant à 3). Signalons en passant que cette notion de fonction mesurable (dont la dénomination n'est d'ailleurs pas heureuse) est très importante et se définit de façon générale dans la situation suivante : $\mathcal{A}$ étant une tribu de parties d'un ensemble X, et $\mathcal{B}$ une tribu de parties d'un ensemble φ, une application f de X dans φ est dite ($\mathcal{A}$, %)-mesurable si l'image réciproque de tout élément de $\mathcal{B}$ est un élément de $\mathcal{A}$. Si une mesure μ a été définie sur $\mathcal{A}$, cela permet d'en définir une ν sur $\mathcal{B}$, en posant, pour tout $B \in \mathcal{B}$:

$$\nu(B) = \mu[f^{-1}(B)].$$

Pour les fonctions bornées, la différence essentielle entre l'intégration de Riemann et celle de Lebesgue réside dans le fait que la première part d'un clan et d'une mesure simplement additive et la seconde d'une tribu et d'une mesure o-additive, fait qui avait été obscurci par certains commentaires déclarant que, dans le cas de fonctions réelles définies sur un segment de R,

Riemann partageait le domaine de la variable et Lebesgue celui de la fonction, ce qui ne correspondait qu'à deux manières d'obtenir des fonctions étagées (que ni l'un ni l'autre n'utilisaient explicitement).

Lebesgue ne se limitait en outre pas aux fonctions bornées ou définies sur des ensembles de mesure finie, mais il étendait son intégrale de telle sorte qu'elle apparaissait finalement comme une forme linéaire positive I, classiquement notée :

$$I(f) = \int_X f\,d\mu,$$

sur l'espace $\mathcal{L}^1$, ou, si l'on veut préciser, $\mathcal{L}^1(X, \mathcal{B}, \mu)$, des fonctions intégrables qui est un espace vectoriel réticulé, avec les propriétés suivantes :

(CI') $f \in \mathcal{L}^1$ est équivalent à $|f| \in \mathcal{L}^1$.

(b') Sur $\mathcal{L}^1$, $I(|f|)$ est une semi-norme. La relation :

$$I(|f-g|) = 0$$

est équivalente au fait que f et g prennent les mêmes valeurs sauf aux points d'un ensemble de mesure nulle, ce que l'on traduit en disant qu'elles sont égales *presque partout*. L'égalité presque partout est une relation d'équivalence, et l'espace quotient est un espace vectoriel normé, classiquement noté $L^1(X, \mathcal{B}, \mu)$ qui a la remarquable propriété d'être complet,

(c') Si f_n est une suite croissante de fonctions intégrables, telles que la suite IV;,) converge, la suite f_n converge simplement presque partout (c'est-à-dire sauf aux points d'un ensemble de mesure nulle) vers une fonction intégrable f telle que :

$$I(f) = \lim_n I(f_n).$$

(d') *Théorème de la convergence dominée.* Si une suite de fonctions intégrables f_n converge presque partout vers une fonction f et s'il existe une fonction intégrable g telle que, pour tout $n \in N$, on ait $|f_n| \leqslant g$, alors f est intégrable et :

$$I(f) = \lim_n I(f_n).$$

4. L'intégrale comme forme linéaire

Le fait que l'intégrale est une forme linéaire sur un espace vectoriel de fonctions est si fondamental qu'il peut en constituer une définition ; cependant cette importance n'était pas encore perçue au moment où Lebesgue créait son intégrale. Un des résultats qui contribua le plus à dégager le rôle de cette notion fut le théorème de F. Riesz, déjà cité, sur l'identité entre les intégrales de Stieltjes des fonctions continues réelles définies sur un segment $[a, b]$ et les formes linéaires continues sur l'espace de Banach que constituent ces fonctions. Les idées de Riesz furent étendues par J. Radon, dont le nom est désormais associé aux formes linéaires continues sur l'espace V des fonctions continues à support compact définies sur un espace localement compact X, l'espace V étant muni de la topologie de la convergence compacte ; l'hypothèse de continuité que l'on impose ici à une forme linéaire m sur V s'exprime par le fait que, pour tout compact K de X, il existe une constante M(K) telle que :

$$m(f) \leqslant M(K)\|f\|_\infty,$$

où :

$$\|f\|_\infty = \sup_{x \in X} |f(x)|$$

pour toute fonction $f \in V$ nulle en dehors de K. L'aspect linéaire a paru tellement important aux mathématiciens contemporains, et en particulier à Bourbaki, qu'ils

utilisent le terme de mesure *de Radon* pour désigner non plus des fonctions σ-additives d'ensembles, mais les formes linéaires décrites ci-dessus.

Un autre pas en direction de la linéarisation fut accompli par l'Américain P. J. Daniell, qui exposa une théorie de l'intégration comme méthode de prolongement d'une forme linéaire positive présentant une « continuité » convenable.

Une théorie de l'intégration est d'abord l'étude du prolongement d'une forme linéaire, continue en un certain sens, sur un espace vectoriel de fonctions à un espace vectoriel plus vaste. Le cœur de la question est que cet espace plus vaste se présente naturellement sous deux formes différentes – comme complété d'un espace vectoriel topologique dont les éléments sont a priori des classes d'équivalence de suites de Cauchy (cf. espaces MÉTRIQUES, espaces vectoriels NORMÉS), ou comme quotient, relativement à une relation d'égalité presque partout, d'espaces de fonctions) – dont il s'agit de montrer qu'elles sont isomorphes.

Voici le schéma d'une telle théorie (selon M. H. Stone).

(a'') Le départ est un espace vectoriel V réticulé d'applications d'un ensemble X dans R, contenant $\inf(1, f)$ s'il contient f, et une forme linéaire positive I sur V telle que :

$$|f| \leqslant \sum_{n\in\mathbf{N}} |f_n| \Rightarrow \mathrm{I}(|f|) \leqslant \sum_{n\in\mathbf{N}} \mathrm{I}(|f_n|).$$

(b'') Cela étant, on associe à toute application f de X dans R la quantité $\mathrm{N}(f)$ (réelle positive ou égale à $+\infty$) définie par :

$$\mathrm{N}(f) = \inf\Big(\sum_{n\in\mathbf{N}} \mathrm{I}(|f_n|)\Big)$$

où la borne inférieure est prise par rapport à l'ensemble des suites (f_n) de fonctions de V telles que :

$$|f| \leqslant \sum_{n\in\mathbf{N}} |f_n| \,;$$

on a alors :

$$\mathrm{N}\Big(\sum_{i\in\mathbf{N}} f_i\Big) \leqslant \sum_{i\in\mathbf{N}} \mathrm{N}(f_i)\,;$$

on se restreint alors à l'ensemble G des fonctions f pour lesquelles $\mathrm{N}(f)$ est *fini*. L'ensemble G est un espace vectoriel qui contient V et sur lequel N est une semi-norme.

Le résultat essentiel est que G est *complet* pour cette semi-norme.

(c'') On considère alors l'adhérence $\overline{\mathrm{V}}$ de V dans G qui est un espace complet pour la semi-norme N ; on peut prolonger de manière unique, par continuité, I en une forme linéaire $\overline{\mathrm{I}}$ continue (relativement à N) sur $\overline{\mathrm{V}}$. Puisque $\overline{\mathrm{I}}$ s'annule sur le sous-espace des fonctions telles que $\mathrm{N}(f) = 0$, on peut définir une forme J sur l'espace quotient $\mathcal{V}$ de $\overline{\mathrm{V}}$ pour la relation d'équivalence $\mathrm{N}(f-g) = 0$, et $\mathcal{V}$ est un espace de Banach. Si, comme c'est souvent le cas, $\mathrm{N}(f) = 0$ entraîne $f = 0$ si $f \in \mathrm{V}$, alors V est isomorphe à un sous-espace de $\mathcal{V}$ et on peut considérer J comme le prolongement de I de V à $\mathcal{V}$.

$\overline{\mathrm{I}}$ et $\overline{\mathrm{V}}$ ont toutes les propriétés signalées plus haut pour l'intégrale de Lebesgue et les espaces $\mathcal{L}^1(X, \mathcal{B}, \mu)$; $\overline{\mathrm{V}}$ est l'espace $\mathcal{L}^1(X, \mathcal{B}, \mu)$ pour la tribu $\mathcal{B}$ engendrée par les ensembles E dont les fonctions caractéristiques appartiennent à $\overline{\mathrm{V}}$, avec $\mu(E) = \overline{\mathrm{I}}(\chi_E)$. L'espace $\mathcal{V}$ correspond à $L^1(X, \mathcal{B}, \mu)$ et J est la forme déduite de l'intégrale par le passage au quotient.

Une situation extrêmement importante déjà signalée, mais qui entre dans ce cadre

général, est le cas où V est l'espace des fonctions réelles continues à support compact définies sur un espace localement compact et où 1 est une forme linéaire continue sur V muni de la topologie de la convergence compacte, c'est-à-dire le cas des mesures de Radon, qui s'expriment comme différences de deux formes positives.

Enfin signalons, sans donner le moindre détail, qu'il est possible de définir, en en conservant la linéarité et une continuité convenablement définie, des intégrales à valeurs non plus seulement réelles, mais dans C, dans R" et plus généralement dans les espaces vectoriels topologiques localement convexes.

Espace L^p

Aux espaces $\mathcal{L}^1$ et $L^1(X, \mathcal{B}, \mu)$ peuvent être associés d'autres espaces remarquables, dont une des définitions peut être présentée comme suit : Appelant mesurables les fonctions réelles, qui sont $(\mathcal{B}, B)$-mesurables, où B est la tribu borélienne de R, l'espace $\mathcal{L}^p$ pour $p \geqslant 1$, est constitué des fonctions mesurables telles que $|f|^p \in \mathcal{L}^1$. C'est un espace vectoriel sur lequel N_p définie par :

$$N_p(f) = \left(\int_X |f|^p d\mu\right)^{1/p}$$

est une semi-norme. La relation $N_p(f) = 0$ est équivalente à l'égalité presque partout, et l'espace quotient relativement à cette relation, $L^p(X, \mathcal{B}, \mu)$, est normé et complet.

De plus, si $p > 1$. le dual topologique de L^p est isomorphe à L^q si :

$$\frac{1}{P} + \frac{1}{4} = 1,$$

et on a l'inégalité dite de Hölder, pour $f \in L^p$ et $g \in L^q$:

$$\left|\int_X fg\, d\mu\right| \leqslant N_p(f) N_q(g).$$

Pour $p = 2$, on a $q = 2$ et l'espace L^2 est isomorphe à son dual topologique. L'inégalité de Hölder prend la forme particulière appelée inégalité de Schwarz :

$$\left|\int_X fg\, d\mu\right| \leqslant N_2(f) N_2(g);$$

$\int_X fg\, d\mu$ ne dépend que des classes de f et de g et permet de définir un produit scalaire sur L^2 qui apparaît alors comme un espace de Hilbert et joue un rôle extrêmement important en analyse harmonique et en physique quantique.

On désigne par $\mathcal{L}^\infty$ l'espace des fonctions mesurables bornées presque partout. C'est un espace vectoriel sur lequel on peut définir une semi-norme $N_\infty(f)$, égale au plus petit réel positif k tel que l'ensemble :

$$\{x \in X \mid f(x) > k\}$$

soit de mesure nulle. L'espace quotient, relativement à l'égalité presque partout, est noté L^∞. II est normé et complet. Il est isomorphe au dual topologique de L^1, mais il n'y a pas cette fois réciprocité ; le dual de L^∞ contient au moins un sous-espace isomorphe à L^1, mais le contient strictement.

Mesure de Haar

La longueur, l'aire, le volume sont des mesures de Radon invariantes par les translations de R, $\mathbf{R}^2$, $\mathbf{R}^3$.

Un résultat très général et très important, dû à Haar, généralise cette situation : Soit G un groupe topologique localement compact, dont l'opération est notée multiplicativement. Si s est un élément du groupe, à toute fonction continue à support compactfon peut associer la fonction f_s, également continue à support compact. définie par :

$$f_s(x) = f(s^{-1}x)$$

Le théorème de Haar affirme qu'il existe alors une mesure de Radon sur G unique (à un facteur près) *invariante à gauche,* c'est-à-dire telle que, pour tout $s \in G$ et pour toute fonction f,

$$I(f_s) = I(f).$$

5. Intégration et dérivation

Un très célèbre théorème d'analyse classique énonce que, si f est une fonction *continue* réelle définie sur $[a, b]$, l'application :

$$x \mapsto \int_a^x f(t)\,dt$$

est dérivable et admet $f(x)$ pour dérivée au point x.

En vertu de ce théorème, intégration et dérivation sont souvent présentées comme des « opérations inverses » l'une de l'autre.

En réalité, la recherche des primitives (ce qui est vraiment l'inversion de la dérivation) et l'intégration ne coïncident nullement, car les fonctions intégrables ne sont pas toutes des fonctions dérivées, et les fonctions dérivées ne sont pas **toutes** intégrables.

Le problème de la recherche des primitives **de** la fonction dérivée la plus générale a été résolu par A. Denjoy dans sa belle et difficile théorie de la totalisation.

Mais le problème peut être présenté autrement : partant d'un espace mesuré $(X, \mathcal{B}, \mu)$ et d'une fonction intégrable f, on peut définir une mesure ν sur $\mathcal{B}$, en prenant pour valeur $\nu(A)$ de la mesure d'un élément A de $\mathcal{B}$ l'intégrale de $f\chi_A$, où χ_A est la fonction caractéristique de A. Réciproquement, ν étant une mesure sur $\mathcal{B}$, existe-t-il une fonction intégrable f telle que, pour tout $A \in \mathcal{B}$, on ait :

$$\nu(A) = \int f\chi_A\,d\mu\ ?$$

La réponse est fournie par le théorème de Radon-Nikodym (d'ailleurs énoncée et démontrée par Lebesgue dans le cas où μ est la mesure de Lebesgue sur R), que nous ne donnerons pas dans sa plus grande généralité : Si $X \in \mathcal{B}$ et $\mu(X) < +\infty$, la condition nécessaire et suffisante pour que ν puisse s'exprimer par :

$$\nu(A) = \int f\chi_A\,d\mu,$$

est que, pour tout ensemble B de $\mathcal{B}$,

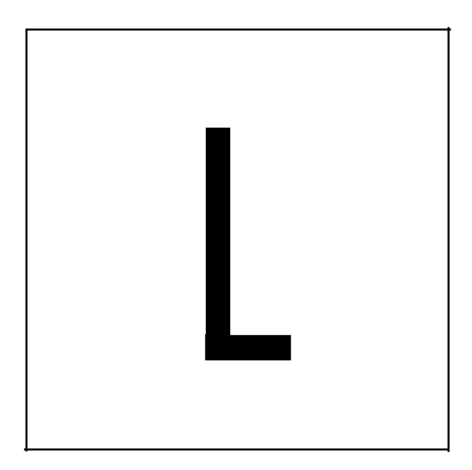
$$\mu(B) = 0 \Rightarrow \nu(B) = 0.$$

La fonction f n'est évidemment déterminée que presque partout, mais est unique à cela près, et peut être considérée comme une densité de la mesure ν par rapport à la mesure μ : c'est, en fait, la meilleure manière et la plus générale de concevoir la notion de densité.

ANDRÉ REVUZ

Bibliographie

R. COUTY & J. AZRA, *Analyse*, Armand Colin, Paris, 5' éd. 1980 / P. DEHEUVELS, *L'Intégrale*, coll. Que sais-je ?. P.U.F., 1986 / P. HALMOS, *Measure Theory*, Springer-Verlag, New York, 1991 / H. LEBESGUE. *Leçons sur l'intégration*, Gauthier-Villars. Paris, 1928, reprod. en fac-sim. J. Gabay, Paris, 1989 / *La Mesure des grandeurs*, Blanchard, Paris, 1975 / J. NEVEU, *Bases mathématiques du calcul des probabilités*, Masson, Paris, 2' éd. 1970 / J. NEVEU & M. MÉTIVIER, *Théorie de la mesure et de l'intégration*, École polytechnique, Palaiseau, 1983 / K. VO KHAC. *Mesure, intégration, convolution et analyse* de Fourier, Marketing, Paris, 1984.



LIE GROUPES DE → GROUPES · Groupes de Lie

LIMITE NOTION DE

La notion de limite fait son apparition dans un ouvrage du mathématicien anglais B. Robins intitulé A *Discourse Concerning the Nature and Certainty of Sir Isaac Newton's Method of Fluxions und Prime and Ultimate Ratios* (1135) ; c'est une réponse aux critiques formulées par le philosophe G. Berkeley à l'encontre du calcul infinitésimal dans son célèbre pamphlet *The Analyst* (1734). Robins essaie de préciser et de clarifier l'expression un peu obscure de Newton « premières et dernières raisons », en parlant de *limites* vers quoi tendent, sans jamais les atteindre, des rapports de quantités variables ; il a dû soutenir une controverse contre son compatriote J. Jurin, newtonien orthodoxe et sourcilleux, pour qui les premières et dernières raisons étaient effectivement atteintes (à l'instant de naissance ou d'évanouissement).

C. Maclaurin, dans son *Treatise of Fluxions* (1742), présenté lui aussi comme une réponse à Berkeley, reprend l'interprétation des « premières et dernières raisons » de Newton en termes de limites ; cependant il fonde le calcul infinitésimal sur la notion de fluxion (vitesse instantanée) et non sur celle de limite. Au contraire, d'Alembert, dans l'article « Différentiel » de *L'Encyclopédie*, vol. IV, 1754, présente la notion de limite comme la « vraie métaphysique du calcul différentiel » : il y définit le rapport différentiel dy/dx comme la limite du rapport des accroissements finis de y et de x lorsque ces accroissements tendent vers 0, et il insiste sur le fait que l'on ne doit pas séparer les « différentielles » dy et dx. Comme pour ses prédécesseurs Robins et Maclaurin, le langage de D'Alembert est entièrement géométrique, et la notion de limite n'est pas très clairement définie : on dit simplement que le rapport considéré peut devenir aussi proche que l'on veut de sa limite, ou encore qu'une « grandeur est la limite d'une autre grandeur, quand la seconde peut s'approcher de la première plus près qu'une quantité donnée, si petite qu'on puisse supposer, sans pourtant que la grandeur qui s'approche puisse jamais surpasser la grandeur dont elle s'approche, en sorte que la différence d'une pareille quantité à sa limite est absolument inassignable » (on remarque que, pour d'Alembert, la limite est approchée d'un seul côté). Cependant, d'Alembert prend soin d'établir l'unicité de la limite. Il n'a jamais mis en œuvre son programme de construction du calcul différentiel à partir de la notion de limite : dans tous ses écrits scientifiques, il utilise le langage des infi-

niments petits, langage commun aux mathématiciens continentaux du XVIIIe siècle.

Quelques successeurs de D'Alembert ont donné des exposés du calcul infinitésimal fondés sur la notion de limite. On peut citer A. G. Kästner, *Anfangsgründe der Analysis des Unendlichen* (1761), ouvrage assez maladroit qui comporte des incohérences ; S. L'Huillier, *Exposition élémentaire des principes des calculs supérieurs* (1787), primé par l'Académie de Berlin, où les limites sont présentées comme une interprétation de la « méthode d'exhaustion » des mathématiciens grecs : sa définition de limite n'est pas plus claire que celle de D'Alembert, et toujours en langage géométrique, « Étant donné une quantité variable, toujours plus petite ou plus grande qu'une quantité donnée ; mais qui peut différer de celle-ci de moins qu'une quantité arbitraire, si petite soit-elle ; cette quantité constante est appelée la limite en grandeur ou en petitesse de la quantité variable ». Le *Traité de calcul différentiel et intégral* de Lacroix (1797), qui a connu de nombreuses rééditions et a été traduit en anglais, est aussi fondé sur la notion de limite, et il a sans doute beaucoup fait pour populariser cette notion.

La mise en œuvre de la notion de limite au XVIIIe siècle se heurtait à un certain nombre d'obstacles : le langage géométrique ne fournissait pas un domaine numérique homogène où développer la théorie, et la notion générale de fonction n'était pas encore assimilée. Il était donc difficile de concevoir clairement comment une grandeur ou un rapport variable tendaient vers leurs limites : des objections du genre de celle de Zénon d'Elée pouvaient être opposées, ce qui faisait dire à Lagrange que la notion de limite paraissait soulever des difficultés métaphysiques. Le concept de limite s'est progressivement clarifié au XIXe siècle : dès 1800, C. F. Gauss avait une conception extrêmement claire de la limite d'une suite de nombres réels (a_n), puisqu'il la définit (dans un travail inédit *Notions fondamentales sur la théorie des suites)* comme la valeur commune à lim sup a_n et lim inf a_n lorsque ces deux limites extrêmes, qui sont définies de manière précise, coïncident. A. L. Cauchy a imposé la notion de limite à la base du calcul infinitésimal ; la définition qu'il en donne est encore un peu vague : « Lorsque les valeurs successivement attribuées à une même variable s'approchent indéfiniment d'une valeur fixe, de manière à finir par en différer aussi peu que l'on voudra, cette dernière est appelée la *limite* de toutes les autres » (résumé des « leçons » données à l'École royale polytechnique sur le calcul infinitésimal, 1823) ; mais il introduit une notation lim pour la limite, et il montre sur des exemples numériques comment se comportent les limites.

La définition très précise de limite que l'on donne encore dans les cours remonte à Weierstrass, promoteur du « style des epsilons ». Pour que la théorie soit entièrement claire, il ne manque alors qu'une théorie satisfaisante des nombres réels, qui permettrait d'établir l'existence d'une borne supérieure pour une partie non vide majorée et de démontrer le critère de Cauchy, admis jusqu'alors comme une évidence ; diverses théories des nombres réels ont été élaborées vers 1860-1870 (Dedekind, Weierstrass, Méray, Cantor).

La notion de limite a été étendue hors du cadre numérique par la topologie générale au XXe siècle ; dans les espaces métriques, on peut tout ramener à la définition de la limite d'une suite de points, qui est formellement identique à la définition de la limite d'une suite de nombres, mais dans les espa-

ces plus généraux, les suites ne suffisent plus : on a d'abord utilisé une sorte de généralisation des suites, avec un ensemble d'indices non dénombrable (convergence à la Moore-Smith), puis la notion de *filtre* introduite par H. Cartan (1937) ; de cette dernière notion est dérivée celle d'ultrafiltre qui a fourni un puissant moyen de construction et de démonstration en topologie générale et en logique.

CHRISTIAN HOUZEL

LINÉAIRE & MULTILINÉAIRE
ALGÈBRE

L'algèbre linéaire sur un corps commutatif, telle qu'on la trouvera présentée ici, s'est progressivement dégagée, au cours du XIX[e] siècle et au début du XX[e], de la théorie des équations linéaires (systèmes de n équations linéaires à p inconnues, équations différentielles et intégrales linéaires) et de la géométrie (calcul vectoriel dans les espaces affines, transformations des espaces projectifs, dualité pour les sous-variétés linéaires et les quadriques, structure même de la géométrie). L'algèbre multilinéaire sur un corps commutatif a pris naissance dans la théorie des invariants et dans la partie de la géométrie différentielle consacrée au calcul tensoriel. Plus récemment, on a développé l'algèbre linéaire sur un anneau afin d'appliquer les méthodes de l'algèbre linéaire sur les corps à la théorie des groupes abéliens, considérés comme Z-modules, à la théorie des entiers algébriques sur un anneau commutatif unitaire, considérés comme éléments d'un module sur cet anneau, à la représentation linéaire d'un groupe dans un espace vectoriel, considéré comme module sur l'algèbre de ce groupe, et à l'étude des formes quadratiques sur Z. Enfin, ces dernières années ont été introduites l'algèbre homologique et, plus généralement, la théorie des catégories abéliennes, permettant d'appliquer la théorie des modules à des domaines où elle semblait inopérante (théorie des fibrés vectoriels et des faisceaux).

On trouvera un aperçu historique plus complet dans l'article ALGÈBRE. D'autre part, on trouvera des détails sur les applications de l'algèbre linéaire dans de nombreux articles tels que GROUPES - Groupes classiques et géométrie, Groupes de Lie et théorie des NOMBRES - Nombres algébriques. Bien entendu, la liste précédente n'est pas exhaustive : on pourrait, à la limite, affirmer que l'algèbre linéaire a envahi tous les domaines des mathématiques. À titre d'exemple, on consultera les applications à la théorie des équations algébriques et à l'analyse fonctionnelle (cf. équations aux DÉRIVÉES PARTIELLES, DISTRIBUTIONS, algèbres NORMÉES, espaces vectoriels NORMÉS).

Dans le présent article sera d'abord exposée la théorie des espaces vectoriels sur un corps commutatif, indépendamment de la notion de dimension. L'explicitation des résultats obtenus lorsque les espaces vectoriels sont de dimension finie et munis de bases fait l'objet du paragraphe consacré au calcul matriciel. Dans cette partie, l'exposé reste élémentaire, et la plupart des théorèmes sont accompagnés de démonstrations. Suivent quelques indications sur l'algèbre tensorielle et la théorie des modules.

En ce qui concerne la réduction des endomorphismes et la théorie des formes quadratiques, on se reportera aux articles : théorie SPECTRALE et formes QUADRATIQUES.



1. Espaces vectoriels et applications linéaires

Espaces vectoriels

Soit K un corps commutatif, On appelle *espace vectoriel* sur K, ou encore K-espace vectoriel. un ensemble E muni de deux lois de composition : une loi interne, application de E X E dans E, notée $(x, y) \mapsto x + y$ et une loi externe, application de K $\times$ E dans K, notée $(a, x) \mapsto a \cdot x$, ou encore $(a, x) \mapsto \alpha x$; ces deux lois satisfaisant aux conditions suivantes :

(a) L'ensemble E, muni de l'addition, est un groupe commutatif.

(b) Pour tout couple (a, β) d'éléments de K et pour tout élément x de E :

$$\alpha . (\beta . x) = (\alpha\beta) . x,$$

et, pour tout élément x de E, $1 . x = x$.

(c) Pour tout couple (a, β) d'éléments de K et pour tout couple (x, y) d'éléments de E :

$$(\alpha + \beta) . x = \alpha . x + \beta . x,$$
$$\alpha . (x + y) = \alpha . x + \alpha . y.$$

Les éléments de E sont souvent appelés *vecteur.~,* les éléments de K *scalaires.*

Applications linéaires

Soit E et F deux espaces vectoriels sur un même corps commutatif K. On dit qu'une application U de E dans F est K-linéaire ou, plus simplement, *linéaire* si, pour tout couple (x, y) d'éléments de E et pour tout couple (a, β) de scalaires :

$$U(\alpha x + \beta y) = \alpha U(x) + \beta U(y).$$

On dit aussi que U est un morphisme d'espaces vectoriels.

Soit E, F et G trois espaces vectoriels sur K. Pour toute application linéaire U de E dans F et pour toute application linéaire V de F dans G, l'application composée $V \circ U$ est linéaire.

On dit qu'une application linéaire U de E dans F est un *isomorphisme* de E sur F s'il existe une application linéaire V de F dans E telle que :

$$V \circ U = I_E \quad \text{et} \quad U \circ V = I_F.$$

Une application linéaire de E dans lui-même s'appelle *endomorphisme* de E, et un isomorphisme de E sur lui-même *automorphisme de* E.

Voici quelques exemples d'espaces vectoriels et d'applications linéaires :

1. Soit n un entier naturel non nul. L'ensemble K^n des suites de n éléments de K, muni des deux lois définies par les formules :

$$(\xi_1, \xi_2, ..., \xi_n) + (\eta_1, \eta_2, ..., \eta_n) = (\xi_1 + \eta_1, \xi_2 + \eta_2, ..., \xi_n + \eta_n),$$
$$\alpha(\xi_1, \xi_2, ..., \xi_n) = (\alpha\xi_1, \alpha\xi_2, ..., \alpha\xi_n),$$

est un espace vectoriel sur K.

2. Soit A un ensemble non vide et F un espace vectoriel sur K. L'ensemble F^A, noté encore $\mathcal{F}(A, F)$, des applications de A dans F, muni des deux lois définies par les formules :

$$(f + g)(x) = f(x) + g(x),$$
$$(\alpha f)(x) = \alpha f(x),$$

est un espace vectoriel sur K.

3. Soit $(F_i)_{i \in I}$ une famille d'espaces vectoriels sur un même corps commutatif K. L'ensemble produit :

$$\prod_{i \in I} F_i,$$

muni des deux lois suivantes :

$$(x_i)_{i \in I} + (y_i)_{i \in I} = (x_i + y_i)_{i \in I},$$
$$\alpha(x_i)_{i \in I} = (\alpha x_i)_{i \in I}.$$

est un espace vectoriel sur K, appelé espace vectoriel produit de la famille $(F_i)_{i \in I}$. (Lorsque tous les espaces vectoriels F_i sont égaux à un même espace vectoriel F, l'espace produit n'est autre que F^I.) Pour tout élément j de I, le projecteur canonique de l'espace produit sur F,, qui à toute famille $(x_i)_{i \in I}$ associe le vecteur x_j, est une application linéaire surjective.

4. Soit E un espace vectoriel sur K. On appelle *forme linéaire* sur E une application linéaire de E dans K, le corps K étant considéré comme espace vectoriel sur lui-même.

Par exemple, soit E l'espace vectoriel des fonctions continues sur l'intervalle [0, 1] à valeurs complexes. L'application qui à tout élémentfde E associe le scalaire :

$$\int_0^1 f(x)\,dx$$

est une forme linéaire sur E.

Sous-espaces vectoriels

Soit E un espace vectoriel sur K, $(x_i)_{i \in I}$ une famille de vecteurs de E, $(\alpha_i)_{i \in I}$ une famille de scalaires dont le support J est fini. (On appelle support de $(\alpha_i)_{i \in I}$ l'ensemble J des éléments i de I tels que $\alpha_i \neq 0$.) Pour toute partie finie H de I contenant J :

$$\sum_{i \in H} \alpha_i x_i = \sum_{i \in J} \alpha_i x_i$$

Cette somme se note encore :

$$\sum_{i \in I} \alpha_i x_i.$$

Cette convention permet de poser la définition suivante : On dit qu'un vecteur x de E est *combinaison linéaire* des vecteurs x_i s'il existe une famille $(\alpha_i)_{i \in I}$ de scalaires à support fini telle que :

$$x = \sum_{i \in I} \alpha_i x_i.$$

Par exemple, soit E l'espace vectoriel des fonctions à valeurs réelles définies sur **R**, et $(f_n)_{n \in \mathbf{N}}$ la famille des fonctions monomiales $f_n : x \mapsto x^n$. Les combinaisons linéaires de ces fonctions ne sont autres que les fonctions polynomiales. En revanche, la fonction exponentielle $x \mapsto e^x$ n'est pas combinaison linéaire des fonctions f_n.

Soit E un espace vectoriel sur K. On dit qu'une partie E' de E est un *sous-espace vectoriel* de E si E' est stable pour les deux lois de E et si, munie des lois induites, E' est un espace vectoriel sur K.

Pour qu'une partie non vide E' de E soit un sous-espace vectoriel de E, il faut et il suffit que, pour tout couple (x, y) d'éléments de E' et pour tout couple (α, β) de scalaires, le vecteur $\alpha x + \beta y$ appartienne à E'.

L'intersection d'une famille de sous-espaces vectoriels de E est encore un sous-espace vectoriel de E. Il en découle que, pour toute partie A de E, l'ensemble des sous-espaces vectoriels de E contenant A possède un plus petit élément (au sens de la relation d'inclusion), à savoir l'intersection de tous les sous-espaces vectoriels contenant A. Ce sous-espace vectoriel, dit *engendré* par A, est encore l'ensemble des combinaisons linéaires d'éléments de A, lorsque A est non vide. Voici quelques exemples.

Le sous-espace vectoriel engendré par un vecteur x est noté Kx; c'est en effet l'ensemble des vecteurs de E de la forme αx, où α appartient à K.

Soit E un espace vectoriel sur K et $(E_i)_{i \in I}$ une famille de sous-espaces vectoriels de E. Le sous-espace vectoriel de E engendré par la réunion des sous-espaces vectoriels E_i est constitué des vecteurs de E de la forme :

$$x = \sum_{i \in I} x_i,$$

où, pour tout élément i de I, x_i appartient à E_i, et où la famille $(x_i)_{i \in I}$ est à support fini. Ce sous-espace vectoriel s'appelle aussi *somme* des sous-espaces vectoriels E_i, et se note :

$$\sum_{i \in I} E_i.$$

Dans le cas particulier où $I = \{1, 2\}$, la somme des sous-espaces vectoriels E_1 et E_2 se note $E_1 + E_2$.

Espaces vectoriels d'applications linéaires

Soit E et F deux espaces vectoriels sur K. L'ensemble des applications linéaires de E dans F est un sous-espace vectoriel, noté $\mathcal{L}(E, F)$, de l'espace vectoriel $\mathcal{F}(E, F)$ des applications de E dans F.

Soit E, F et G trois espaces vectoriels sur K. L'application $V \mapsto V \circ U$ est une application linéaire de $\mathcal{L}(F, G)$ dans $\mathcal{L}(E, G)$, et l'application $U \mapsto V \circ U$ une application linéaire de C(E, F) dans $\mathcal{L}(E, G)$.

En particulier, l'ensemble des endomorphismes d'un espace vectoriel E, muni des trois lois de composition :

$$(U, V) \mapsto U + V,$$
$$(U, V) \mapsto V \circ U,$$
$$(\alpha, U) \mapsto \alpha U,$$

est une algèbre associative unitaire, notée $\mathcal{L}(E)$ (cf. ANNEAUX ET ALGÈBRES). Le produit $V \circ U$ se note encore VU.

Les automorphismes de E constituent un groupe multiplicatif, appelé groupe linéaire de E, et noté GL(E) ; c'est le groupe multiplicatif des éléments inversibles de l'anneau unitaire C(E).

Factorisation des applications linéaires

Soit E et F deux espaces vectoriels sur K, et U une application linéaire de E dans F. L'image d'un sous-espace vectoriel de E par U est un sous-espace vectoriel de F. En particulier, l'image de E par U est un sous-espace vectoriel de F, appelé aussi *image* de U, et noté Im(U). De même, l'image réciproque d'un sous-espace vectoriel de F par U est un sous-espace vectoriel de E. En particulier, l'image réciproque du sous-espace vectoriel réduit au vecteur nul de F est un sous-espace vectoriel de E, appelé *noyau* de U, et noté Ker(U). Pour que U soit injective, il faut et il suffit que son noyau soit réduit au vecteur nul de E.

Théorème 1 (théorème de factorisation). Soit E, F et G : trois espaces vectoriels sur K.

1. Soit U une application linéaire surjective de E sur F. Pour toute application linéaire V de E dans G telle que Ker(V) contienne Ker(U), il existe une application linéaire W et une seule de F dans G telle que $V = W \circ U$. Plus précisément, l'application $W \mapsto W \circ U$ est un isomorphisme de l'espace vectoriel $\mathcal{L}(F, G)$ sur le sous-espace vectoriel de C(E, G) constitué des applications linéaires dont le noyau contient celui de U.

2. Soit U une application linéaire injective de F dans G. Pour toute application linéaire V de E dans G telle que Im(V) soit contenue dans Im(U), il existe une application linéaire W et une seule de E dans F telle que $V = U \circ W$. Plus précisément. l'application W-W $\circ$ U est un isomorphisme de l'espace vectoriel $\mathcal{L}(E, F)$ sur le sous-espace vectoriel de C(E, G) constitué des applications linéaires dont l'image est contenue dans celle de U.

Espaces vectoriels quotients

Soit E' un sous-espace vectoriel d'un espace vectoriel E. La relation binaire dans E définie par les couples (x, y) tels que $x - y$ appartienne à E' est compatible avec les lois de E. Muni des lois quotients, l'ensemble

quotient est un espace vectoriel sur K, appelé espace vectoriel quotient de E par E', et noté E/E'. L'application canonique φ de E sur E/E' est linéaire, et son noyau **est** E'.

Le théorème de factorisation montre aussitôt que le couple $(E/E', \varphi)$ possède la propriété universelle suivante :

Pour tout couple (F, U) constitué d'un espace vectoriel F sur K et d'une application linéaire U de E dans F dont le noyau contient E', il existe une application linéaire $\tilde{U}$ et une seule de E/E' dans F telle que $U = \tilde{U} \circ \varphi$. Plus précisément, l'application $V \mapsto V \circ \varphi$ est un isomorphisme de l'espace vectoriel $\mathcal{L}(E/E', F)$ sur le sous-espace vectoriel de C(E, F) constitué des applications linéaires de E dans F dont le noyau contient E'.

Voici une conséquence immédiate de la propriété universelle des espaces vectoriels quotients : Soit E et F deux espaces vectoriels sur K, soit U une application linéaire de E dans F, soit φ l'application canonique de E sur $E/\mathrm{Ker}(U)$, soit V l'unique application de $E/\mathrm{Ker}(U)$ dans $\mathrm{Im}(U)$ telle que U(x) = (V o <p)(x), pour tout vecteur x de E, et soit i l'injection canonique de $\mathrm{Im}(U)$ dans F. Alors V est un isomorphisme de $E/\mathrm{Ker}(U)$ sur Im(U), et :

$$u = i \circ V \circ \varphi,$$

formule de décomposition canonique de U, qui ramène en quelque sorte l'étude de U à celles de i, de φ et de V.

Dualité

Soit E un espace vectoriel sur K. L'espace vectoriel C(E, K) des formes linéaires sur E s'appelle espace vectoriel *dual* de E, et se note E*. L'application de E* X E dans K, qui au couple (y^*, x) associe le scalaire $y^*(x)$, est une forme bilinéaire, dite canonique, et encore notée :

$$(y^*, x) \mapsto \langle y^*, x \rangle.$$

Le dual de l'espace vectoriel E*, c'est-à-dire l'espace vectoriel des formes linéaires sur E*, s'appelle *bidual* de E, et se note E**. Pour éviter des confusions, nous noterons :

$$(z^{**}, y^*) \mapsto \langle\langle z^{**}, y^* \rangle\rangle$$

la forme bilinéaire canonique sur E** X E*.

Étant donné un vecteur x de E, l'application de E* dans K, qui **à toute** forme linéaire y^* sur E associe le scalaire $\langle y^*, x \rangle$, est une forme linéaire sur E* ; c'est donc un élément de E**. L'application χ, qui associe au vecteur x cet élément de E**, est une application linéaire de E dans E**, dite canonique ; elle est définie par la relation :

$$\langle\langle \chi(x), y^* \rangle\rangle = \langle y^{**}, x \rangle.$$

On dit qu'un vecteur x de E et une forme linéaire y* sur E sont *orthogonaux* si $\langle y^*, x \rangle = 0$. On dit qu'une partie A de E et une partie B de E* sont orthogonales si, pour tout élément x de A et pour tout élément y^* de B, x et y^* sont orthogonaux. L'ensemble des éléments de E* orthogonaux à un sous-espace vectoriel F de E est un sous-espace vectoriel de E*, appelé orthogonal de F, et noté $F^{\perp}$. De même, l'ensemble des vecteurs de E orthogonaux à un sous-espace vectoriel G de E* est **un** sous-espace vectoriel de E, appelé orthogonal de G, et noté G'.

Théorème 2. Soit E et F deux espaces vectoriels sur K, soit E* et F* leurs duaux, et U une application linéaire de E dans F. Il existe une application linéaire de F* dans E* et une seule, appelée *transposée* de U et notée ${}^tU$. telle que. pour tout élément x de E et pour tout élément y^* de F*,

$$(1) \qquad \langle {}^tU(y^*), x \rangle = \langle y^*, U(x) \rangle.$$

L'aplication ${}^tU$ n'est autre que l'application $y^* \mapsto y^* \circ U$. La relation (1) s'appelle identité fondamentale de la transposition.

L'application $U \mapsto {}^tU$ est une application linéaire de $\mathcal{L}(E, F)$ dans $\mathcal{L}(F^*, E^*)$, appelée transposition.

La transposée de l'application identique de E n'est autre que l'application identique de son dual :

$$ {}^t I_E = I_{E^*} $$

Soit E, F et G trois espaces vectoriels sur K, soit U une application linéaire de E dans F, et V une application linéaire de F dans G. Alors la transposée de $V \circ U$ est égale à ${}^tU \circ {}^tV$. En particulier, si G = E, et si U est inversible à gauche (resp. à droite). ${}^tU$ est inversible à droite (resp. à gauche). Plus particulièrement encore, si U est un isomorphisme de E sur F, ${}^tU$ est un isomorphisme de F* sur E*, et :

$$ ({}^tU)^{-1} = {}^t(U^{-1}) $$

L'isomorphisme de E* sur F* ainsi défini s'appelle *contragrédient* de U, et se note $\check{U}$. Lorsque F = E, l'application $U \mapsto \check{U}$ est un morphisme du groupe GL(E) dans le groupe GL(E*).

Enfin, pour toute application linéaire U de E dans F, le noyau de ${}^tU$ n'est autre que l'orthogonal dans F* de l'image de U :

$$ \mathrm{Ker}({}^tU) = [\mathrm{Im}(U)]^{\perp} $$

Équations linéaires

Soit E et F deux espaces vectoriels sur K, soit U une application linéaire de E dans F, et b un élément de F. On appelle *équation linéaire* définie par U et b l'équation :

$$ (1) \qquad U(x) = b. $$

Le vecteur b s'appelle second membre de l'équation (1). Lorsque $b = 0$, on dit que l'équation (1) est homogène, ou, par abus de langage, sans second membre. L'équation :

$$ (2) \qquad U(x) = 0 $$

s'appelle équation linéaire homogène associée à l'équation (1).

Voici les propriétés de l'ensemble des solutions d'une équation linéaire.

Si l'équation linéaire (1) est homogène, ses solutions constituent un sous-espace vectoriel de E, à savoir le noyau de U. Dans le cas général, si l'équation (1) admet une solution x_0, on obtient toutes les solutions de cette équation en ajoutant à x_0 une solution quelconque de l'équation homogène associée.

Enfin, pour que l'équation (1) admette une solution et une seule quel que soit le second membre b, il faut et il suffit que l'application linéaire U soit bijective, ou encore que sa transposée ${}^tU$ le soit. Dans ces conditions, l'unique solution de l'équation (1) n'est autre que $U^{-1}(b)$; l'application de F dans E, qui à tout vecteur b associe cette solution, est donc linéaire. On dit aussi que la solution dépend linéairement du second membre.

On voit donc que les notions d'image et de noyau sont fondamentales pour l'étude des équations linéaires.

2. Sommes directes, bases

Sommes directes

Soit $(E_i)_{i\in I}$ une famille d'espaces vectoriels sur K. Dans l'espace vectoriel :

$$ \prod_{i\in I} E_i, $$

l'ensemble des éléments $(x_i)_{i\in I}$ à support fini est un sous-espace vectoriel de cet

espace vectoriel, appelé *somme directe* de la famille $(E_i)_{i\in I}$, et noté :

$$\bigoplus_{i\in I} E_i\,;$$

il coïncide avec l'espace vectoriel produit lorsque l'ensemble I est fini.

Soit, en particulier, E un espace vectoriel sur K, soit $(E_i)_{i\in I}$ une famille de sous-espaces vectoriels de E, et U l'application linéaire de la somme directe de cette famille dans E qui à tout élément $(x_i)_{i\in I}$ associe l'élément :

$$\sum_{i\in I} x_i.$$

Alors l'image de U est la somme :

$$\sum_{i\in I} E_i$$

des sous-espaces vectoriels E_i, et le noyau de U est l'ensemble des éléments $(x_i)_{i\in I}$ tels que :

$$\sum_{i\in I} x_i = 0$$

Ainsi, pour que U soit surjective, il faut et il suffit que :

$$\sum_{i\in I} E_i = E,$$

et, pour que U soit injective, il faut et il suffit que, pour tout $i \in I$:

$$E_i \cap \sum_{j\neq i} E_j = \{0\}.$$

Lorsque ces deux conditions sont réalisées, c'est-à-dire lorsque U est un isomorphisme, il est d'usage d'identifier E et :

$$\bigoplus_{i\in I} E_i,$$

ce qui conduit à dire que E est *somme directe des sous-espaces vectoriels* E_i.

Enfin, pour que E soit somme directe des sous-espaces vectoriels E_i, il faut et il suffit que tout vecteur x de E s'écrive d'une manière et d'une seule sous la forme :

$$x = \sum_{i\in I} x_i,$$

où, pour tout élément i de I, x_i appartient à E_i.

L'intérêt de la notion de somme directe apparaît dans le théorème suivant.

Théorème 3. Soit E et F deux espaces vectoriels sur K, et $(E_j)_{j\in J}$ une famille de sous-espaces vectoriels de E dont E est somme directe.

1. Pour tout élément $(U_j)_{j\in J}$ de :

$$\prod_{j\in J} \mathcal{L}(E_j, F),$$

il existe une application linéaire U et une seule de E dans F telle que, pour tout élément j de J, la restriction de U à E_j soit égale à U_j. À tout vecteur x de E, écrit sous la forme :

$$x = \sum_{j\in J} x_j,$$

où, pour tout $j \in J$, $x_j \in E_j$, l'application U associe le vecteur :

$$\sum_{j\in J} U_j(x_j).$$

2. L'application $(U_j)_{j\in J} \mapsto U$ est un isomorphisme de l'espace vectoriel :

$$\prod_{j\in J} \mathcal{L}(E_j, F)$$

sur l'espace vectoriel :

$$\mathcal{L}\Big(\bigoplus_{j\in J} E_j, F\Big)$$

En particulier, deux applications linéaires de E dans F ayant, pour tout élément j

de J, même restriction au sous-espace vectoriel E, sont égales.

Sous-espaces vectoriels supplémentaires, projecteurs

On dit que deux sous-espaces vectoriels F et G d'un espace vectoriel E sur K sont *supplémentaires* dans E si les trois conditions équivalentes suivantes sont vérifiées :

(a) L'espace vectoriel E est somme directe de F et de G.

(b) Tout vecteur x de E s'écrit d'une manière et d'une seule sous la forme $x = y + z$, où $y \in F$ et $z \in G$.

(c) La réunion de F et de G engendre E, et l'intersection de F et de G est réduite au vecteur nul.

L'application P_F qui associe au vecteur x le vecteur y est un endomorphisme de E, appelé *projecteur* sur F parallèlement à G. Le vecteur y est appelé projection de x sur F parallèlement à G. On définit de même P,.

Le projecteur P_F a pour image F et pour noyau G, et les endomorphismes P_F et P_G satisfont aux relations :

(1) $$P_F P_G = P_G P_F = 0$$

(2) $$P_F^2 = P_F \text{ et } P_G^2 = P_G$$

(3) $$P_F + P_G = 1,.$$

Les seules relations $P_F + P_G = I_E$ et $P_F^2 = P_F$ impliquent les relations (1) à (3). En effet $P_F P_G = P_F(I_E \quad P_F) = P_F - P_F^2 = 0$; de même, $P_G P_F = 0$; enfin :

$$P_G^2 = (I_E - P_F)^2 = I_E - 2P_F + P_F^2 = I_E - P_F = P_G.$$

C'est pourquoi on dit qu'un endomorphisme U de E est un projecteur si $U^2 = U$. L'endomorphisme U est alors le projecteur sur Im(U) parallèlement à Ker(U).

Par exemple, dans l'espace vectoriel $\mathfrak{F}(A, K)$ des applications d'un ensemble A dans K, le sous-espace vectoriel F des applications nulles en un point donné a de A et le sous-espace vectoriel G des applications constantes sont supplémentaires. Le projecteur sur G parallèlement à F est l'application $f \mapsto f(a)$.

De même, dans l'espace vectoriel sur C des fonctions n fois continûment dérivables sur R à valeurs complexes, le sous-espace vectoriel F des fonctions polynomiales de degré inférieur ou égal à n et le sous-espace vectoriel G constitué des fonctions f telles que, pour tout $p \in [0, n]$, $(D^p f)(0) = 0$ sont supplémentaires. Le projecteur sur F parallèlement à G n'est autre que l'application qui à toute fonction associe son développement limité à l'ordre n au point 0.

Soit enfin E un espace vectoriel sur K et $(E_i)_{i \in I}$ une famille de sous-espaces vectoriels de E dont E est somme directe. Pout tout élément i de I, l'application P, qui associe au vecteur x sa i-ième composante x_i est le projecteur sur E_i parallèlement à :

$$\bigoplus_{j \neq i} E_j$$

La famille des projecteurs P, satisfait aux relations suivantes :

- Pour tout couple (i, j) d'éléments distincts de I,

$$P_i P_j = P_j P_i = 0 \text{ ;}$$

Pour tout élément i de I,

$$P_i^2 = P, \text{ ;}$$

- L'application identique de E est égale à la somme des projecteurs P,.

Une famille $(P_i)_{i \in I}$ de projecteurs de E satisfaisant aux trois conditions précédentes s'appelle système de projecteurs.

L'intérêt de la notion de sous-espaces vectoriels supplémentaires apparaît dans le théorème fondamental suivant.

Théorème 4. Soit E et F deux espaces vectoriels sur K, U une application linéaire de E dans F, E' un sous-espace vectoriel de E, et U' la restriction de U à E'. Pour que U' définisse un isomorphisme de E' sur Im(U), il faut et il suffit que E' soit un sous-espace vectoriel supplémentaire de Ker(U).

Soit en particulier E, un sous-espace vectoriel d'un espace vectoriel E sur K. Pour tout sous-espace vectoriel E_2 supplémentaire de E, dans E, la restriction à E_2 de l'application linéaire canonique de E sur E/E_1 est un isomorphisme de E_2 sur E/E_1. (Il suffit d'appliquer le théorème précédent au cas où F = E/E_1, où U est l'application linéaire canonique de E sur E/E_1, et où E' = E_2.)

Soit enfin E_2 et E'_2 deux sous-espaces vectoriels supplémentaires de E, dans E. La restriction à E_2 du projecteur sur E'_2 parallèlement à E, définit un isomorphisme de E_2 sur E'_2. (Il suffit cette fois d'appliquer le théorème précédent au cas où F = E',, où U est le projecteur sur E'_2 parallèlement à E,, et où E' = E_2.)

Ainsi, deux sous-espaces vectoriels supplémentaires d'une même troisième sont canoniquement isomorphes.

Bases

Soit K^I l'espace vectoriel des applications d'un ensemble non vide I dans K. L'ensemble, noté $K^{(I)}$, des applications de I dans K à support fini est un sous-espace vectoriel de K^I ; il est égal à K^I si 1 est fini.

En particulier, prenons pour I l'ensemble N des entiers naturels. Alors K^N est l'espace vectoriel des séries formelles à coefficients dans K, tandis que $K^{(N)}$ est l'espace vectoriel des polynômes à coefficients dans K.

On notera que :

$$K^I = \prod_{i\in I} E_i \quad \text{et} \quad K^{(I)} = \bigoplus_{i\in I} E_i,$$

où, pour tout élément i de I, $E_i = K$. Par suite, l'espace vectoriel $K^{(I)}$ est somme directe des sous-espaces vectoriels Ke_i, où pour tout élément i de 1, e, est l'élément de $K^{(I)}$ défini par les formules $e_i(j) = \delta_{ij}$. Autrement dit, tout élémentfde $K^{(I)}$ s'écrit d'une manière et d'une seule sous la forme :

$$f = \sum_{i\in I} \alpha_i e_i,$$

où α_i n'est autre que $f(i)$.

Théorème 5. Le couple $(K^{(I)}, (e_i)_{i\in I})$ possède la propriété universelle suivante : Pour tout couple (E, $(x_i)_{i\in I}$) constitué d'un espace vectoriel E sur K et d'une famille $(x_i)_{i\in I}$ de vecteurs de E, il existe une application linéaire U et une seule de $K^{(I)}$ dans E telle que, pour tout élément i de 1, $U(e_i) = x_i$. L'application U associe à tout élément $f = (\alpha_i)_{i\in I}$ le vecteur :

$$\sum_{i\in I} \alpha_i x_i.$$

L'image de U est le sous-espace vectoriel de E engendré par les vecteurs x_i, et le noyau de U est l'ensemble des éléments $(\alpha_i)_{i\in I}$ tels que :

$$\sum_{i\in I} \alpha_i x_i = 0.$$

Les éléments de ce noyau sont appelés *relations linéaires* entre les vecteurs x_i ; en particulier, le vecteur nul de $K^{(I)}$ est appelé relation linéaire triviale entre ces vecteurs.

Ainsi, pour que U soit surjective, il faut et il suffit que tout vecteur x de E soit une

combinaison linéaire des vecteurs x_i. On dit alors que la famille $(x_i)_{i\in I}$ est *génératrice.* Pour que U soit injective, il faut et il suffit que toute relation linéaire entre les vecteurs x_i soit triviale. On dit alors que la famille $(x_i)_{i\in I}$ **est** *libre.* Lorsque ces deux conditions sont réalisées, c'est-à-dire lorsque U est un isomorphisme de $K^{(I)}$ sur E, on dit que la famille $(x_i)_{i\in I}$ est une *buse* de E. Cela revient à dire que tout vecteur x de E peut s'écrire d'une manière et d'une seule sous la forme :

$$x = \sum_{i\in I} \alpha_i x_i .$$

La famille $(\alpha_i)_{i\in I}$ s'appelle famille des composantes du vecteur x dans la base $(x_i)_{i\in I}$.

Par exemple. la famille $(e_i)_{i\in I}$ est une base, dite canonique, de $K^{(I)}$. En particulier, lorsque 1 = [1, n], la base canonique de K^n est constituée des n vecteurs suivants :

$$e_1 = (1, 0, 0, \ldots . 0),$$
$$e_2 = (0, 1, 0, \ldots . 0),$$
$$\ldots\ldots\ldots\ldots$$
$$e_n = (0, 0, 0, \ldots . 1).$$

De même, l'espace vectoriel $K[X] = K^{(N)}$ des polynômes à une indéterminée à coefficients dans K a pour base canonique la famille des monômes e,, = X^n, où n parcourt N.

Soit S une partie de E. On dit que S est une partie génératrice, une partie libre ou une partie basique si la famille $(x_s)_{s\in S}$, où, pour tout élément s de S, $x_s = s$, est une famille génératrice, une famille libre ou une base.

Pour qu'une partie à un seul élément x soit libre, il **faut** et il suffit que le vecteur x soit non nul. Lorsqu'une partie $\{x, y\}$ à deux éléments n'est pas libre, on dit aussi que les vecteurs x et y sont colinéaires.

La notion de base permet d'exprimer sous la forme suivante la propriété universelle de l'espace vectoriel $K^{(I)}$ énoncée dans le théorème 5.

Théorème 6. Soit E et F deux espaces vectoriels sur K, et $(e_j)_{j\in J}$ une famille d'éléments de E. Si cette famille est génératrice, deux applications linéaires de E dans F prenant pour toutj $\in$ J la même valeur sur le vecteur e_j sont égales. Si cette famille est une base de E, pour toute famille $(f_j)_{j\in J}$ d'éléments de F, il existe une application linéaire U et une seule de E dans F telle que, pour tout $j \in J$, $U(e_j) = f_j$. À tout vecteur x de E décomposé dans la base B sous la forme :

$$x = \sum_{j\in J} \xi_j e_j ,$$

l'application U associe le vecteur :

$$\sum_{j\in J} \xi_j f_j .$$

Nous pouvons maintenant caractériser les applications linéaires injectives et surjectives à l'aide de la transformée d'une base.

Soit E et F deux espaces vectoriels sur K, soit U une application linéaire de E dans F, et $(e_j)_{j\in J}$ une base de E. Pour que U soit surjective (resp. injective), il faut et il suffit que la famille $(U(e_j))_{j\in J}$ soit génératrice (resp. libre). Pour que U soit bijective, il faut et il suffit que $(U(e_j))_{j\in J}$ soit une base de F.

Lorsque F = K, le théorème 6 se particularise de la manière suivante.

Théorème 7. Soit E un espace vectoriel sur K, et B = $(e_j)_{j\in J}$ une base de E. Pour toute famille $(\alpha_j)_{j\in J}$ de scalaires, il existe une forme linéaire y^* et une seule sur E telle que, pour tout $j \in J$, $< y^*, e_j > = \alpha_j$.

À tout vecteur x de E décomposé dans la base B sous la forme :

$$x = \sum_{j \in J} \xi_j e_j,$$

la forme linéaire y^* associe le scalaire :

$$\sum_{j \in J} \xi_j \alpha_j.$$

De plus, l'application $(\alpha_j)_{j \in J} \mapsto y^*$ est un isomorphisme de l'espace vectoriel K^J sur l'espace vectoriel E*.

Ainsi, le dual de $K^{(J)}$ s'identifie à K^J, l'application bilinéaire canonique étant définie par la formule :

$$\langle (\beta_j)_{j \in J}, (\alpha_j)_{j \in J} \rangle = \sum_{j \in J} \alpha_j \beta_j.$$

En particulier, pour **tout** entier naturel n, l'espace $(K^n)^*$ est canoniquement isomorphe à K^n.

3. Existence de bases

Théorème 8. Soit E un espace vectoriel sur K. soit L une partie libre de E, et S une partie génératrice de E contenant L. Il existe alors une partie basique B de E telle que LCBCS.

Nous allons démontrer ce théorème lorsque la partie S est finie.

Introduisons l'ensemble $\mathcal{E}$ ordonné par inclusion des parties libres T de E telles que L CT C S. L'ensemble $\mathcal{E}$ est non vide, puisque L appartient à $\mathcal{E}$. La partie S étant finie, l'ensemble card(T) des entiers naturels, où T parcourt $\mathcal{E}$, admet un plus grand élément p. Soit B un élément de $\mathcal{E}$ ayant p éléments. Montrons que B convient. Puisque B appartient à $\mathcal{E}$, la partie B est libre, et L C B C S. Il reste donc à prouver que B est génératrice. Supposons en effet par l'absurde que le sous-espace vectoriel E' engendré par B ne soit pas égal à E. Puisque S est génératrice, il existe un élément x de S n'appartenant pas à E', ce qui implique que B' = B ∪ $\{x\}$ **est** encore libre. Ainsi, B' est un élément de $\mathcal{E}$ ayant p + 1 éléments, ce qui contredit la définition de p.

Lorsque S est quelconque, la démonstration est analogue, le principe de récurrence étant remplacé par le théorème de Zorn.

Corollaire 1. Pour toute partie libre L de E, il existe une partie basique B de E contenant L ; pour toute partie génératrice S de E, il existe une partie basique B de E contenue dans S. En particulier, pour tout espace vectoriel E sur K, l'ensemble des bases de E est non vide.

Ce corollaire s'obtient en spécialisant le théorème aux trois cas suivants : S = E, L = ∅, S = E et L = ∅.

Corollaire 2 (théorème de la base incomplète). Pour toute partie libre L de E et pour toute partie génératrice S de E, il existe une partie S' de S telle que L ∩ S' soit vide et que B = L ∪ S' soit une partie basique de E.

Voici l'une des principales conséquences du théorème précédent :

Théorème 9. Tout sous-espace vectoriel E' d'un espace vectoriel E admet un sous-espace vectoriel supplémentaire dans E.

On choisit en effet une partie basique B' de E', que l'on complète en une partie basique B de E. Alors le sous-espace vectoriel engendré par B" = B − B' est un sous-espace vectoriel supplémentaire de E' dans E.

Corollaire 1. Soit E et F deux espaces vectoriels sur K, soit E' un sous-espace vectoriel de E, et U' une application linéaire de E' dans F. Il existe alors une application linéaire U de E dans F prolongeant U'.

Corollaire 2. Soit E et F deux espaces vectoriels sur K, soit F' un sous-espace vectoriel de F, et φ l'application linéaire canonique de F sur F/F'. Pour toute application linéaire U de E dans F/F', il existe une application linéaire V de E dans F telle que $U = \varphi \circ V$.

Corollaire 3. Soit E et F deux espaces vectoriels sur K, et U une application linéaire de E dans F. Pour que U soit surjective, il faut et il **suffit** que U soit inversible à droite, c'est-à-dire qu'il existe une application linéaire V de F dans E telle que $U \circ V = I_F$. Pour que U soit injective, il faut et il suffit que U soit inversible à gauche, c'est-à-dire qu'il existe une application linéaire V de F dans E telle que $V \circ U = I_E$.

En effet. il est évident que. si U est inversible à droite (resp. à gauche), U est surjective (resp. injective). Réciproquement, si U est surjective, U définit un isomorphisme U' d'un supplémentaire E' de Ker(U) sur F ; il suffit de prendre pour V l'application linéaire de F dans E coïncidant avec U'^{-1}. De même, si U est injective, U définit un isomorphisme U' de E sur Im(U) ; il suffit alors de prendre pour V l'application linéaire nulle sur un supplémentaire F' de Im(U), et coïncidant avec U'^{-1} sur Im(U).

4. Espaces vectoriels de dimension finie

Définition

On dit qu'un espace vectoriel E sur K est de dimension finie sur K, ou, plus simplement. de dimension finie, s'il existe une partie basique finie de E. Dans le cas contraire, on dit que E est de dimension infinie.

Pour qu'un espace vectoriel E soit de dimension finie, il faut et il suffit qu'il existe une partie génératrice finie de E, puisque de toute partie génératrice on peut extraire une partie basique.

Théorème 10. Soit E un espace vectoriel de dimension finie sur K, et B une partie basique finie de E ayant n éléments. Alors toute partie libre L de E est finie, et le nombre p d'éléments de L est inférieur ou égal à n. De plus, on peut compléter L en une partie basique de E en lui adjoignant $(n - p)$ éléments convenablement choisis dans B.

Le théorème se démontre en utilisant le lemme d'échange suivant, qui fournit en outre un procédé pratique de complétion de L en une partie basique.

Lemme. Soit $B = (e_1, e_2, ..., e_n)$ une base de E, soit q un entier inférieur ou égal à n, et $L_q = (f_1, f_2, ..., f_q)$ une famille libre de E. On suppose que $B_q = (f_1, f_2, ..., f_{q-1}, e_q, ..., e_n)$ est une base de E. Alors il existe au moins un entier $i \in [q, n]$ tel qu'en substituant f_q à e_i dans B_q on obtienne encore une base de E, notée B_{q+1}.

Il suffit pour cela de décomposer f_q dans la base B_q. Puisque L_q est libre, il existe au moins un entier $i \in [q, n]$ tel que la i-ième composante de f_q soit non nulle. Il est alors immédiat que cet entier i convient.

Corollaire 1. Soit E un espace vectoriel de dimension finie sur K. Toutes les parties basiques de E sont finies, et elles ont le même nombre d'éléments.

Il résulte en effet du théorème 10 que toutes les parties basiques de E sont finies. Soit donc B et B' deux parties basiques de E. ayant respectivement n et n' éléments. Comme B est basique et que B' est libre, $n' \leqslant n$; de même, $n \leqslant n'$, et finalement $n = n'$.

Le cardinal commun a toutes les parties basiques de E s'appelle *dimension* de E sur K, et se note $\dim_K E$, ou, plus simplement, dim E. L'espace vectoriel réduit au vecteur nul est le seul espace vectoriel de dimension 0. Un espace vectoriel de dimension 1 s'appelle une *droite*, un espace vectoriel de dimension 2 s'appelle un *plan*.

Voici quelques exemples :

Soit 1 un ensemble non vide. L'espace vectoriel $K^{(I)}$ est de dimension finie si et seulement si I est fini, et la dimension de $K^{(I)}$ est alors égale à card(I). En particulier, pour tout entier naturel non nul n, K^n est de dimension n.

Soit E et F deux espaces vectoriels sur K non réduits à $\{0\}$, soit B = $(e_j)_{j\in J}$ une base de E, et B' $= (f_i)_{i\in I}$ une base de F. Pour tout élément (i, j) de I X J, désignons par U_{ij} l'unique application linéaire de E dans F telle que, pour tout élément k de J,

$$U_{ij}(e_k) = 0 \text{ si } k \neq j \text{ et } U_{ij}(e_j) = f_i.$$

Les applications linéaires U,, constituent une base de l'espace vectoriel $\mathfrak{L}(E, F)$, dite associée aux bases B et B'. En particulier, si E et F sont de dimension finie, il en est de même de C(E, F), et :

$$\dim C(E, F) = (\dim E) . (\dim F).$$

(Notons que cette formule reste valable si E ou F est réduit à $\{0\}$.)

Plus particulièrement encore, l'espace vectoriel des endomorphismes de E est de dimension finie, et :

$$\dim L:(E) = (\dim E)^2.$$

Soit E un espace vectoriel sur K non réduit à $\{0\}$, et B = $(e_j)_{j\in J}$ une base de E. Pour tout élément i de J, l'unique forme linéaire e_i^* telle que, pour tout élément j de J :

$$\langle e_i^*, e_j\rangle = \delta_{ij}$$

s'appelle i-ième forme linéaire coordonnée. La famille $(e^*_j)_{j\in J}$ des formes linéaires coordonnées est libre ; pour que ce soit une base de E*, il faut et il suffit que E soit de dimension finie. Cette base s'appelle base duale de la base B, et se note B*. Nous voyons ainsi que l'espace vectoriel dual de E est de dimension finie si et seulement si E est de dimension finie, et que, dans ces conditions :

$$\dim E^* = \dim E.$$

Corollaire 2. Pour que deux espaces vectoriels de dimension finie sur K soient isomorphes, il faut et il suffit qu'ils aient même dimension.

Corollaire 3. Soit E un espace vectoriel de dimension finie n sur K, et S une partie génératrice finie de E. Alors le nombre p d'éléments de S est supérieur ou égal à n. De plus, il existe une partie basique de E constituée de n vecteurs convenablement choisis dans S.

Corollaire 4. Soit E un espace vectoriel de dimension finie n sur K. Toute partie libre de E ayant n éléments est une partie basique de E, et toute partie génératrice de E ayant n éléments est une partie basique de E.

Dimension et codimension d'un sous-espace vectoriel

Soit E un espace vectoriel sur K. On dit qu'un sous-espace vectoriel E' de E est de codimension finie dans E si l'espace vectoriel quotient E/E' est de dimension finie. La dimension de E/E' s'appelle alors *codimension* de E' dans E, et se note $\operatorname{codim}_E E'$. Les sous-espaces vectoriels de codimension 1 dans E s'appellent *hyperplans* de E.

Pour qu'un sous-espace vectoriel E' de E soit de codimension finie dans E, il faut et il suffit que E' admette un sous-espace

vectoriel supplémentaire de dimension finie. Alors, pour tout sous-espace vectoriel E” supplémentaire de E’ dans E :

$$\dim E'' = \operatorname{codim}_E E' = \dim E/E'.$$

Théorème II. Soit E et F deux espaces vectoriels sur K, et U une application linéaire de E dans F.

1. Si F est de dimension finie et si U est injective, alors E est de dimension finie, et dim E $\leqslant$ dim F. En particulier, tout sous-espace vectoriel F’ d’un espace vectoriel F de dimension finie est aussi de dimension finie, et dim F’ $\leqslant$ dim F.

2. Si E est de dimension finie et si U est surjective, alors F est de dimension finie, et dim F $\leqslant$ dim E. En particulier, tout sous-espace vectoriel E’ d’un espace vectoriel E de dimension finie est de codimension finie dans E, et :

$$\dim E/E' = \operatorname{codim}_E E' \leqslant \dim E.$$

Soit en effet B une partie basique de E, et B’ une partie basique de E’.

Si U est injective, card(U(B)) = card(B), et U(B) est une partie libre de F. Il résulte du théorème 10 que card(U(B)) $\leqslant$ card(B’). L’assertion 1 en découle, puisque F est de dimension finie. Le cas particulier s’en déduit en prenant pour E un sous-espace vectoriel F’ de F, et pour U l’injection canonique de F’ dans F.

Si U est surjective, U(B) est une partie génératrice de F ; d’autre part, card (U(B)) $\leqslant$ card(B). Il résulte du corollaire 4 du théorème 10 que card(B’) $\leqslant$ card(U(B)). L’assertion 2 en découle, puisque E est de dimension finie. Le cas particulier s’en déduit en prenant pour F l’espace vectoriel quotient E/E’, et pour U l’application linéaire canonique de E sur E/E’.

Corollaire. Soit E un espace vectoriel de dimension finie sur K, et E’ un sous-espace vectoriel de E. Alors la dimension de tous les sous-espaces vectoriels supplémentaires de E’ dans E est égale à dim E − dim E’. De plus :

$$\operatorname{codim}_E E' = \dim E/E' = \dim E - \dim E'.$$

En particulier, pour que E’ = E, il faut et il suffit que dim E’ = dim E. De même, les hyperplans d’un espace vectoriel de dimension n ne sont autres que les sous-espaces vectoriels de dimension $n - 1$.

Rang d’une application linéaire

Soit E et F deux espaces vectoriels sur K, et U une application linéaire de E dans F. On dit que U est de rang fini si l’image de U est un espace vectoriel de dimension finie. La dimension de Im(U) s’appelle alors *rang* de l’application linéaire U, et se note rang(U).

Théorème 12. Pour que l’application linéaire U soit de rang fini, il faut et il suffit que le noyau de U soit de codimension finie dans E. Le rang de U est alors égal à la codimension dans E du noyau de U :

$$(1)\quad \operatorname{rang}(U) = \dim \operatorname{Im}(U) = \operatorname{codim}_E \operatorname{Ker}(U)$$

En effet, l’espace vectoriel quotient E/Ker(U) est isomorphe à l’espace vectoriel Im(U).

Si F est de dimension finie, toute application linéaire U de E dans F est de rang fini, car Im(U), étant un sous-espace vectoriel de F, est de dimension finie.

Si E est de dimension finie, toute application linéaire U de E dans F est de dimension finie, et l’on a la formule de la dimension :

$$(2)\qquad \operatorname{rang}(U) = \dim E - \dim \operatorname{Ker}(U).$$

En effet, U définissant une application linéaire surjective de E sur Im(U), l’espace vectoriel Im(U) est de dimension finie. La

formule (2) est alors une conséquence de la formule (1) et de la suivante :

$$\text{codim}_{\text{,}} \text{Ker(U)} = \dim E - \dim \text{Ker(U)}.$$

Lorsque les espaces vectoriels E et F sont tous deux de dimension finie, et qu'ils ont même dimension n, il est équivalent de dire :

L'application linéaire U est un isomorphisme de E sur F ;
L'application linéaire U est inversible à droite ;
- L'application linéaire U est inversible à gauche ;
L'application linéaire U est bijective ;
- L'application linéaire U est surjective ;
L'application linéaire U est injective ;
- Le rang de U est égal à n.

Dualité en dimension finie

Théorème 13. Soit E un espace vectoriel de dimension finie sur K. Alors l'application linéaire canonique χ de E dans son bidual E** est un isomorphisme.

Soit en effet x un élément du noyau de χ. Alors. pour toute forme linéaire y^* sur E,

$$\langle \chi(x), y^* \rangle = \langle y^*, x \rangle = 0.$$

Choisissons une base $B = (e_j)_{1 \leqslant j \leqslant n}$ de E. En prenant successivement pour y^* les n formes linéaires coordonnées e_i^*, nous voyons que toutes les composantes de x sont nulles, et donc que $x = 0$, ce qui montre que l'application linéaire χ est injective. Comme :

$$\dim E^{**} = \dim E^* = \dim E,$$

la formule de la dimension permet d'en déduire que χ est un isomorphisme de E sur E**.

Il découle de ce théorème que l'application qui à toute base B de E associe sa base duale B* est une bijection de l'ensemble des bases de E sur l'ensemble des bases de E*.

On peut maintenant préciser les propriétés de l'orthogonalité en dimension finie.

Théorème 14. Soit E un espace vectoriel de dimension finie sur K, et E* son dual.

Pour tout sous-espace vectoriel F de E :

(1) $$\dim F + \dim F^{\perp} = \dim E.$$

Pour tout sous-espace vectoriel G de E* :

(2) $$\dim G + \dim G' = \dim E.$$

De plus :

(3) $$(F^{\perp})' = F \quad \text{et} \quad (G')^{\perp} = G.$$

Considérons en effet une base $(e_1, e_2, \ldots . e_p)$ de F, et complétons-la en une base $(e_1, e_2, \ldots . e_n)$ de E ; soit $(e_1^*, e_2^*, \ldots . e_n^*)$ sa base duale. Pour établir la formule (1), il suffit de prouver que $(e_{p+1}^*, e_{p+2}^*, \ldots . e_n^*)$ est une base de $F^{\perp}$. Les formes linéaires coordonnees e_{p+1}^*, e_{p+2}^*, $\ldots . e_n^*$, appartiennent évidemment à $F^{\perp}$, car elles sont orthogonales aux vecteurs $e_1, e_2, \ldots . e_p$, lesquels engendrent F ; elles sont linéairement indépendantes, car elles font partie d'une base ; enfin, il reste à montrer que tout élément y^* de $F^{\perp}$ est combinaison linéaire de ces éléments. Écrivons pour cela y^* sous la forme :

$$y^* = \sum_{i=1}^{n} \eta_i e_i^*.$$

Puisque, pour tout élément i de $[1, p]$, $\langle y^*, e_i \rangle = 0$, nous voyons que $\eta_1 = \eta_2 = \ldots = \eta_p = 0$, ce qui achève la démonstration. La formule (2) s'établit de manière analogue, en tenant compte du fait que toute base de E* est la base duale d'une base de E. Enfin, la formule (3) se déduit des relations $F \subset (F^{\perp})'$, $G \subset (G')^{\perp}$, $\dim F = \dim (F^{\perp})'$ et $\dim G = \dim (G')^{\perp}$, ces deux dernières égalités découlant des formules (1) et (2).

Théorème 15. Soit E et F deux espaces vectoriels de dimension finie sur K, soit U une application linéaire de E dans F, et ‘U sa transposée.

Les applications linéaires U et ‘U ont même rang :

$$\text{rang}(U) = \text{rang}({}^t U).$$

En effet, la formule de la dimension appliquée à ${}^t U$ montre que :

$$\text{rang}({}^t U) = \dim F^* - \dim \text{Ker}({}^t U) = \dim F - \dim \text{Ker}({}^t U).$$

D’autre part :

$$\text{rang}(U) = \text{dii } \text{Im}(U) = \dim F - \dim[\text{Im}(U)]^{\perp}.$$

La formule annoncée résulte alors du fait que le noyau de ${}^t U$ n’est autre que l’orthogonal de l’image de U.

5. Matrices

Matrices et applications linéaires

Soit E et F deux espaces vectoriels sur K non réduits à $\{0\}$, de dimensions respectives p et n, soit B = $(e_1, e_2, \ldots .. e_p)$ une base de E, soit B’ = $(f_1, f_2, \ldots . f_n)$ une base de F et U une application linéaire de E dans F. Pour tout élémentj de $[1, p]$, le vecteur U(e,) se décompose d’une manière et d’une seule dans la base B’ sous la forme :

$$(1) \qquad U(e_j) = \sum_{i=1}^{n} \alpha_{ij} f_i.$$

Ainsi, à toute application linéaire U de E dans F nous pouvons associer une famille (α_{ij}) d’éléments de K. Réciproquement, pour toute famille (a,) d’éléments de K, où $(i, j) \in [1, n] \times [1, p]$, il existe une application linéaire U et une seule de E dans F satisfaisant aux conditions (1).

Nous sommes donc amené à introduire les définitions suivantes, utiles pour les calculs explicites concernant les applications linéaires : Soit K un corps commutatif, n et p deux entiers naturels non nuls. On appelle matrice à n lignes et p colonnes à éléments dans K toute famille :

$$M = (\alpha_{ij}), \quad (i, j) \in [1, n] \times [1, p]$$

d’éléments de K. Il est d’usage de disposer les éléments d’une matrice dans les cases d’un **tableau** rectangulaire à n lignes et p colonnes, encadré de deux parenthèses (ou parfois de deux crochets) :

$$M = \begin{pmatrix} \alpha_{11} & \alpha_{12} \ldots & \alpha_{1j} \ldots & \alpha_{1p} \\ \alpha_{21} & a,, & \alpha_{2j} & \alpha_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{i1} & \alpha_{i2} & \alpha_{ij} & \alpha_{ip} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1} & \alpha_{n2} & a, & \alpha_{np} \end{pmatrix}$$

L’indice i s’appelle indice de ligne, l’indice j, indice de colonne. Pour tout élément i de $[1, n]$, la suite $(\alpha_{ij})_{1 \leqslant j \leqslant p}$ s’appelle i-ième ligne de M ; pour tout élément j de $[1, p]$, la suite $(\alpha_{ij})_{1 \leqslant i \leqslant n}$ s’appelle j-ième colonne de M.

Le vecteur de $\mathbf{K}^p$ dont les composantes constituent la i-ième ligne de M s’appelle i-ième ligne de M ; le vecteur de $\mathbf{K}^n$ dont les composantes constituent la j-ième colonne de M s’appellej-ièmevecteur colonne **de** M.

Lorsque $n = 1$, on dit que M est une matrice ligne ; lorsque $p = 1$, on dit que M est une matrice colonne.

L’ensemble des matrices à n lignes et p colonnes à éléments dans K, se note $\mathbf{M}_{n,p}(K)$. Lorsque $n = p$, on dit que A4 est une matrice carrée d’ordre ***n.*** L’ensemble des matrices carrées d’ordre ***n à*** éléments dans K se note M,(K).

Soit, plus généralement, A, I et J trois ensembles, I et J étant finis. On appelle

matrice de type (I, J) à éléments dans A **toute** famille :

$$M=(a,,), \quad (i,j)\in \mathrm{I}\times \mathrm{J}$$

d'éléments de A. Lorsque 1, ou J, est vide, on dit que M est la matrice vide.

Reprenons maintenant le problème initial : la matrice $M = (\alpha_{ij})$ définie par la formule (1) est dite associée à l'application linéaire U dans les bases B et B', et notée $M_{\mathrm{B,B'}}(\mathrm{U})$. La matrice M a pour j-ième colonne la famille des composantes dans la base B' de l'image par U du j-ième vecteur de la base B. L'application qui à toute application linéaire U de E dans F associe la matrice $M_{\mathrm{B,B'}}(\mathrm{U})$ est une bijection de $\mathcal{L}(\mathrm{E, F})$ sur $\mathbf{M}_{n,p}(\mathrm{K})$.

En particulier, lorsque E = F, U est un endomorphisme de E. La matrice $M_{\mathrm{B,B'}}(\mathrm{U})$ est une matrice carrée, appelée matrice associée à l'endomorphisme U dans la base B, et notée plus simplement $M_{\mathrm{B}}(\mathrm{U})$.

Toute matrice peut être considérée comme une matrice associée à une application linéaire : pour tout élément M de M,,,,(K), il existe une application linéaire et une seule de l'espace vectoriel K^p dans l'espace vectoriel K^n dont la matrice associée dans les bases canoniques de ces espaces vectoriels soit M. Cette application linéaire s'appelle *application linéaire de* K^p *dans* K^n *canoniquement associée à M.*

Opérations sur les matrices

La bijection canonique φ de M,,,,(K) sur $\mathcal{L}(\mathrm{K}^p, \mathrm{K}^n)$ ainsi introduite conduit aux définitions qui suivent.

Somme de deux matrices. On appelle somme de deux éléments $M = (\alpha_{ij})$ et M' = (a',,) de $\mathbf{M}_{n,p}(\mathrm{K})$, et on note $M + M'$, l'élément (β_{ij}) de $\mathbf{M}_{n,p}(\mathrm{K})$ défini par les relations :

$$\beta_{ij} = a,, + a',,$$

Produit d'une matrice par un scalaire. On appelle produit d'un élément A4 = (α_{ij}) de $\mathbf{M}_{n,p}(\mathrm{K})$ par un scalaire λ, et on note AM, l'élément (β_{ij}) de $\mathbf{M}_{n,p}(\mathrm{K})$ défini par les relations :

$$\beta_{ij} = \lambda\alpha_{ij}.$$

Muni de ces deux lois, l'ensemble $\mathbf{M}_{n,p}(\mathrm{K})$ est un espace vectoriel de dimension np sur K, et φ est un isomorphisme de l'espace vectoriel $\mathbf{M}_{n,p}(\mathrm{K})$ sur l'espace vectoriel $\mathcal{L}(\mathrm{K}^p, \mathrm{K}^n)$.

Plus généralement, soit **E** et F deux espaces vectoriels sur K, de dimensions respectives p **et** n, soit B une base de E, et B' une base de F. La bijection $\mathrm{U} \mapsto M_{\mathrm{B,B'}}(\mathrm{U})$ est un isomorphisme de l'espace vectoriel L(E, F) sur l'espace vectoriel $\mathbf{M}_{n,p}(\mathrm{K})$.

Produit de deux matrices. Soit m, n et p trois entiers naturels non nuls, soit $M = (\alpha_{ij})$ un élément de M,,,,(K), et N = (β_{hi}) un élément de M,,,,,(K). On appelle produit des matrices Met N, et on note NM, l'élément (γ_{hj}) de M,,,,,(K) défini par les relations :

$$\gamma_{hj} = \sum_{i=1}^{n} \beta_{hi}\,\alpha_{ij}$$

On obtient donc l'élément γ_{hj} à l'intersection de la h-ième ligne et de la j-ième colonne de NM en prenant la h-ième ligne de N, la j-ième colonne de M, et en ajoutant les produits des éléments de même indice (règle de multiplication « ligne par colonne »).

Grâce à la bijection canonique de $\mathbf{M}_{n,p}(\mathrm{K})$ sur $\mathcal{L}(\mathrm{K}^p, \mathrm{K}^n)$, on voit aussitôt que, pour tout couple (M, M') d'éléments de M,,,,(K), pour tout couple (N, N') d'éléments de M,,,,,(K) et pour tout couple (A, μ) de scalaires :

$$(N + N')M = NM + N'M,$$
$$N(M + M') = NM + NM',$$
$$(\mu N)(\lambda M) = (\mu\lambda)NM.$$

Pour tout élément M de $\mathbf{M}_{n,p}(\mathrm{K})$, pour tout élément N de M,,,,,,(K) et pour tout élément P de M,,,,,(K),

$$(PN)M = P(NM).$$

Soit enfin E, F et G trois espaces vectoriels sur K, de dimensions respectives p, n et m, soit B une base de E, soit B' une base de F, et B" une base de G. Pour toute application linéaire U de E dans F et pour toute application linéaire V de F dans G :

$$M_{\mathrm{B,B''}}(\mathrm{V} \circ \mathrm{U}) = M_{\mathrm{B',B''}}(\mathrm{V}) \,.\, M_{\mathrm{B,B'}}(\mathrm{U}).$$

Algèbre des matrices carrées d'ordre n. Muni des trois opérations précédentes, l'ensemble M,,(K) est une algèbre associative unitaire. L'élément unité est la matrice :

$$I_n = (\delta_{ij}).$$

Soit E un espace vectoriel sur K, et B une base de E. L'application $\mathrm{U} \mapsto M_{\mathrm{B}}(\mathrm{U})$ est un isomorphisme de l'algèbre unitaire Ç_(E) sur l'algèbre unitaire M,,(K). Dans cet isomorphisme, le groupe linéaire GL(E) a pour image le groupe, noté $\mathbf{GL}_n(\mathrm{K})$, ou encore $\mathbf{GL}(n, \mathrm{K})$, des matrices carrées d'ordre n inversibles. En particulier, si U est un automorphisme de E, $M_{\mathrm{B}}(\mathrm{U})$ admet pour inverse la matrice associée à U^{-1} dans la base B :

$$[M_{\mathrm{B}}(\mathrm{U})]^{-1} = M_{\mathrm{B}}(\mathrm{U}^{-1}).$$

Transposée d'une matrice. Soit M un élément de $\mathbf{M}_{n,p}(\mathrm{K})$. On appelle transposée de M, et on note ${}^tM$, l'élément de M,,,,(K) dont les colonnes sont les lignes de M.

Soit U l'application linéaire de K^p dans K^n canoniquement associée à M ; alors ${}^tM$ n'est autre que la matrice associée à ${}^t\mathrm{U}$ dans les bases canoniques de $(\mathrm{K}^n)^*$ et de $(\mathrm{K}^p)^*$. Par suite, l'application $M \mapsto {}^tM$ est un isomorphisme de l'espace vectoriel $\mathbf{M}_{n,p}(\mathrm{K})$ sur l'espace vectoriel $\mathbf{M}_{p,n}(\mathrm{K})$.

Pour tout élément M de $\mathbf{M}_{n,p}(\mathrm{K})$ et pour tout élément N de M,,,,(K) :

$${}^t(NM) = {}^tM\,{}^tN$$

Pour tout élément M de $\mathbf{M}_{n,p}(\mathrm{K})$:

$${}^t({}^tM) = M.$$

Soit enfin E et F deux espaces vectoriels sur K, de dimensions respectives p et n, soit B une base de E et B' une base de F, soit B* et B'* leurs bases duales. Pour toute application linéaire U de E dans F :

$$M_{\mathrm{B'^*,B^*}}({}^t\mathrm{U}) = {}^tM_{\mathrm{B,B'}}(\mathrm{U}).$$

Changement de base

Soit E un espace vectoriel de dimension p sur K. Considérons deux bases de E :

$$\mathrm{B}_1 = (e_1, e_2, \dots, e_p), \qquad \mathrm{B}_2 = (e'_1, e'_2, \dots, e'_p),$$

appelées respectivement ancienne et nouvelle base. Pour tout élément j de $[1, p]$, le vecteur e'_j se décompose dans la base B_1 sous la forme :

$$e'_j = \sum_{i=1}^{p} \alpha_{ij} e_i.$$

L'élément $P = (\alpha_{ij})$ de M,,(K) s'appelle matrice de passage de la base B_1 à la base B_2 ; ses colonnes sont constituées des composantes dans l'ancienne base des nouveaux vecteurs de base. La matrice P n'est autre que la matrice $M_{\mathrm{B}_2,\mathrm{B}_1}(\mathrm{I}_{\mathrm{E}})$ associée à l'application identique de E **dans** les bases B_2 et B_1. Il s'ensuit que la matrice P est inversible, et que la matrice de passage de B_2 à B_1 n'est autre que P^{-1}.

Soit E et F deux espaces vectoriels non réduits à $\{0\}$ de dimension finie sur K. soit B, et B_2 deux bases de E et B'_1 et B'_2 deux bases de F, soit P la matrice de passage de B, à B_2 et Q la matrice de passage de B'_1

à B'_2. Pour toute application linéaire U de E dans F, les matrices associées à U dans les bases B, et B',, d'une part, et dans les bases B_2 et B'_2, d'autre part, sont liées par la relation :

$$M_{B_2,B'_2}(U) = Q^{-1}M_{B_1,B'_1}(U)P$$

En particulier, pour tout endomorphisme U de E :

$$M_{B_2}(U) = P^{-1}M_{B_1}(U)P.$$

Rang d'une matrice

Soit M un élément de $\mathbf{M}_{n,p}(K)$. On appelle rang de la matrice M, et on note rang(M) le rang de l'application linéaire de K^p dans K^n canoniquement associée à M.

Plus généralement, soit E et F deux espaces vectoriels de dimension finie sur K non réduits à $\{0\}$, soit B une base de E et B' une base de F. Pour toute application linéaire U de E dans F :

$$\text{rang } M_{B,B'}(U) = \text{rang}(U).$$

De la relation entre le rang d'une application linéaire et celui de sa transposée, on déduit aussitôt que, pour tout élément A4 de $\mathbf{M}_{n,p}(K)$:

$$\text{rang}(M) = \text{rang}({}^tM).$$

Il en découle que le rang de M est égal au rang de ses vecteurs colonnes, ou encore au rang de ses vecteurs lignes.

Les caractérisations des applications linéaires inversibles conduisent à des caractérisations des matrices carrées inversibles.

Soit A4 un élément de M,,(K). Il est équivalent de dire :

La matrice A4 est inversible ;

- La matrice M est inversible à droite ;

La matrice A4 est inversible à gauche ;

- La matrice 'A4 est inversible ;

Le rang de M est égal à n ;

- Le rang des vecteurs colonnes de M est égal à n ;

Le rang des vecteurs lignes de M est égal à n.

Matrices équivalentes

Soit M_1 et M_2 deux éléments de M,,,,(K). On dit que les matrices M_1 et M_2 sont équivalentes s'il existe deux matrices carrées inversibles P et Q d'ordres respectifs p et n à éléments dans K telles que :

$$M_2 = QM_1P.$$

Soit Y un entier naturel. Pour qu'une matrice Mde $\mathbf{M}_{n,p}(K)$ soit de rang Y, il faut et il suffit que M soit équivalente à la matrice $J_r = (\alpha_{ij})$, où $\alpha_{ij} = 1$ si $i \in [1,$ Y] et où $\alpha_{ij} = 0$ dans les autres cas.

Il en découle qu'une condition nécessaire et suffisante pour que deux éléments de $\mathbf{M}_{n,p}(K)$ soient équivalents est qu'ils aient même rang.

On appelle *opérations élémentaires* les applications $M \mapsto M'$ de $\mathbf{M}_{n,p}(K)$ dans lui-même de l'un des types suivants :

(a) La matrice M' se déduit de la matrice M par permutation de deux colonnes, ou de deux lignes.

(b) La matrice M' se déduit de la matrice M par multiplication d'une colonne, ou d'une ligne, par un scalaire non nul.

(c) La matrice M' se déduit de la matrice M par addition à un vecteur colonne (resp. à un vecteur ligne) du produit d'un autre vecteur colonne (resp. d'un autre vecteur ligne) par un scalaire.

Pour que deux matrices **M,** et M_2 soient équivalentes, il faut et il suffit que l'on puisse transformer **M,** en M_2 par une suite finie d'opérations élémentaires.

La théorie des opérations élémentaires permet en outre de calculer le rang d'une matrice, son déterminant et, lorsqu'elle existe, la matrice inverse.

Systèmes d'équations linéaires

Soit n et p deux entiers naturels non nuls, U une application linéaire de K^p dans K^n et $M = (a,)$ la matrice associée, soit $a_1, a_2, \dots a_p$ les vecteurs colonnes de cette matrice et $a'_1, a'_2, \dots a'_n$ ses vecteurs lignes, soit enfin $b = (\beta_i)_{1 \leqslant i \leqslant n}$ un élément de K^n. On désigne par $x = (\xi_j)_{1 \leqslant j \leqslant p}$ un élément de K^n. L'équation $U(x) = b$ équivaut au système de n équations linéaires à p inconnues suivant :

$$\begin{cases} \alpha_{11}\xi_1 + \alpha_{12}\xi_2 + \\ \qquad + \alpha_{1j}\xi_j + \dots + \alpha_{1p}\xi_p = \beta_1 \\ \alpha_{21}\xi_1 + \alpha_{22}\alpha_2 + \\ \qquad + \alpha_{2j}\xi_j + + \alpha_{2p}\xi_p = \beta_2 \\ \dots\dots\dots\dots \\ \alpha_{i1}\xi_1 + \alpha_{i2}\xi_2 + \\ \qquad + \alpha_{ij}\xi_j + + \alpha_{ip}\xi_p = \beta_i \\ \dots\dots\dots\dots \\ \alpha_{n1}\xi_1 + \alpha_{n2}\xi_2 + \\ \qquad + \alpha_{nj}\xi_j + + \alpha_{np}\xi_p = \beta_n \\ \dots\dots\dots\dots \end{cases}$$

La résolution de ce système peut s'interpréter vectoriellement comme la recherche des suites $(\xi_1, \xi_2, \dots \xi_p)$ de scalaires telles que :

$$\xi_1 a_1 + \xi_2 a_2 + + \xi_j a_j + + \xi_p a_p = b,$$

ou encore comme la recherche des vecteurs x de K^p tels que :

$$\begin{cases} \langle a'^*_1, x \rangle = \beta_1 \\ \langle a'^*_2, x \rangle = \beta_2 \\ \dots\dots\dots \\ \langle a'^*_i, x \rangle = \beta_i \\ \dots\dots\dots \\ \langle a'^*_n, x \rangle = \beta_n \end{cases}$$

c'est-à-dire des vecteurs x sur lesquels les formes linéaires $a'^*_1, a'^*_2, \dots a'^*_n$ prennent des valeurs données $\beta_1, \beta_2, \dots \beta_n$.

On appelle ***rung*** de ce système le rang de M, c'est-à-dire le rang de U. En appliquant les propriétés du rang des matrices, on obtient les résultats suivants :

Unicité des solutions. Il est équivalent de dire :

- Pour tout élément b de K^n, l'équation $U(x) = b$ a au plus une solution ;
- Les vecteurs colonnes $a_1, a_2, \dots a_p$ sont linéairement indépendants dans K^n ;

Les vecteurs lignes $a'^*_1, a'^*_2, \dots a'^*_n$ engendrent $(K^p)^*$;

Le rang r de M est égal à p.

Existence des solutions quel que soit le second membre. Il est équivalent de dire :

- Pour tout élément b de K^n, l'équation $U(x)$ a au moins une solution ;
- Les vecteurs colonnes $a_1, a_2, \dots a'_p$ engendrent K^n ;

Les vecteurs lignes $a'^*_1, a'^*_2, \dots a'^*_n$ sont linéairement indépendants dans $(K^p)^*$;

Le rang r de M est égal à n.

Existence d'une solution, le second membre étant donné. Soit b un élément de K^n. Il est équivalent de dire :

L'équation $U(x) = b$ a au moins une solution ;

- Le second membre b appartient au sous-espace vectoriel de K^n engendré par $a_1, a_2, \dots, a_p$;

Le second membre b est orthogonal à $\operatorname{Ker}({}^tU)$;

- Toute relation linéaire vérifiée par les vecteurs lignes l'est aussi par les composantes de b.

6. Applications multilinéaires

Définitions

Soit p un entier naturel non nul, soit $E_1, E_2, \dots, E_p$ et F des espaces vectoriels sur K. On dit qu'une application S de $E_1 \times E_2 \times \dots \times E_p$ dans F est *multilinéaire,* si, pour tout élément j de $[1, p]$, toute application partielle S_j de E_j dans F est linéaire.

Lorsque F = K, on dit que S est une forme multilinéaire.

Soit E un espace vectoriel sur K. Les applications multilinéaires de E^p dans F s'appellent applications p-linéaires sur E à valeurs dans F, et les formes multilinéaires, formes p-linéaires.

Voici quelques cas particuliers :

Soit S une application p-linéaire sur E à valeurs dans F. On dit que S est *alternée* si, pour toute suite $(x_1, x_2, \ldots, x_p)$ de vecteurs de E contenant deux vecteurs égaux :

$$S(x_1, x_2, \ldots, x_p) = 0.$$

On dit que S est *symétrique* si, pour toute permutation σ de [1, p] et pour toute suite $(x_1, x_2, \ldots, x_p)$ de vecteurs de E :

$$S(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(p)}) = S(x_1, x_2, \ldots, x_p).$$

On dit que S est *antisymétrique* si, dans les mêmes conditions :

$$S(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(p)}) = \varepsilon(\sigma) S(x_1, x_2, \ldots, x_p),$$

où $\varepsilon(\sigma)$ désigne la signature de la permutation σ.

Toute application p-linéaire alternée est antisymétrique ; la réciproque devient vraie si la caractéristique du corps K est différente de 2.

Pour qu'une application p-linéaire S soit alternée, il faut et il suffit que, pour toute suite $(x_1, x_2, \ldots, x_p)$ de vecteurs de E contenant deux vecteurs consécutifs x_i et x_{i+1} égaux :

$$S(x_1, x_2, \ldots, x_p) = 0.$$

Si S est alternée, on ne change pas la valeur de S sur une suite de p vecteurs de E en ajoutant à l'un de ces vecteurs une combinaison linéaire des autres. En particulier, si l'un des vecteurs $x_1, x_2, \ldots, x_p$ est une combinaison linéaire des autres, $S(x_1, x_2, \ldots, x_p) = 0$.

Les applications p-linéaires sur E à valeurs dans F constituent un sous-espace vectoriel, noté $\mathcal{M}_p(E, F)$, de l'espace vectoriel $\mathcal{F}(E^p, F)$ des applications de E^p dans F. Lorsque $p = 1$, $\mathcal{M}_p(E, F)$ n'est autre que $\mathcal{L}(E, F)$. Les applications p-linéaires symétriques et les applications p-linéaires alternées constituent des sous-espaces vectoriels de $\mathcal{M}_p(E, F)$, notés respectivement $\mathcal{S}_p(E, F)$ et $\mathcal{A}_p(E, F)$. Enfin, lorsque F = K, ces divers espaces vectoriels se notent plus simplement $\mathcal{M}_p(E)$, $\mathcal{S}_p(E)$ et $\mathcal{A}_p(E)$.

Extension d'une application linéaire

Voici une généralisation de la transposition : Soit E, E' et F trois espaces vectoriels sur K, et U une application linéaire de E' dans E. Pour toute application p-linéaire S sur E à valeurs dans F, l'application S_U de E'^p dans F définie par la formule :

$$S_U(x_1, x_2, \ldots, x_p) = S(U(x_1), U(x_2), \ldots, U(x_p))$$

est une application p-linéaire sur E' à valeurs dans F. L'application U_p qui à tout élément S de $\mathcal{M}_p(E, F)$ associe S_U est une application linéaire de $\mathcal{M}_p(E, F)$ dans $\mathcal{M}_p(E', F)$, et l'image de U_p de $\mathcal{S}_p(E, F)$ (resp. de $\mathcal{A}_p(E, F)$) est contenue dans $\mathcal{S}_p(E', F)$ (resp. dans $\mathcal{A}_p(E', F)$). Soit enfin E" un espace vectoriel sur K, et V une application linéaire de E" dans E'. Alors :

$$(V \circ U)_p = U_p \circ V_p.$$

Lorsque U est l'homothétie de rapport α dans l'espace vectoriel E, l'application U_p n'est autre que l'homothétie de rapport α^p dans l'espace vectoriel $\mathcal{M}_p(E, F)$.

Voici maintenant une méthode générale de construction d'applications p-linéaires symétriques, ou alternées : Pour toute application p-linéaire S de E dans F, les

applications M(S) et A(S) de E^p dans F, définies pour les formules :

$$M(S)(x_1, x_2, ..., x_p) = \sum_{\sigma \in \Sigma_p} S(x_{\sigma(1)}, x_{\sigma(2)}, ..., x_{\sigma(p)}),$$

$$A(S)(x_1, x_2, ..., x_p) = \sum_{\sigma \in \Sigma_p} \varepsilon(\sigma) S(x_{\sigma(1)}, x_{\sigma(2)}, ..., x_{\sigma(p)}),$$

où Σ_p désigne le groupe symétrique de degré p, sont respectivement symétrique et alternée. De plus, l'application M : S-M(S) de $\mathcal{M}_p(E, F)$ dans $\mathcal{S}_p(E, F)$ est linéaire, et l'application A : S-A(S) de $\mathcal{M}_p(E, F)$ dans $\mathcal{A}_p(E, F)$ est linéaire ; elles se dénomment opérateurs de symétrisation et d'antisymétrisation.

Formes multilinéaires

Lorsque F = K, nous allons étudier la structure de $\mathcal{M}_p(E)$, de S,(E) et de $\mathcal{A}_p(E)$.

Nous allons d'abord construire des formes p-linéaires à l'aide de formes linéaires.

Soit $(y_1^*, y_2^*, , y_p^*)$ une suite de p formes linéaires sur E. L'application f de E^p dans K définie par la formule :

$$f(x_1, x_2, ..., x_p) = \langle y_1^*, x_1 \rangle \langle y_2^*, x_2 \rangle \langle y_p^*, x_p \rangle$$

est une forme p-linéaire sur E, appelée *produit tensoriel* des formes linéaires y_1^*, y_2^*, ... y_p^* et notée $y_1^* \otimes y_2^* \otimes \otimes y_p^*$. Les applications $M(f)$ et $A(f)$, symétrisée et antisymétrisée de f, s'appellent respectivement *produit symétrique* et *produit extérieur* des formes linéaires y_1^*, y_2^*, ..., y_p^* et se notent $y_1^* \cdot y_2^* \cdot \cdot y_p^*$ et $y_1^* \wedge y_2^* \wedge \wedge$ y;,.

L'application qui à $(y_1^*, y_2^*, ... y_p^*)$ associe $y_1^* \otimes y_2^* \otimes \otimes y_p^*$ (resp. $y_1^* \cdot y_2^* \cdot ... \cdot y_p^*$, resp. y_1^* A y_2^* A A y_p^*) est une application y-linéaire (resp. p-linéaire symétrique, resp. p-linéaire alternée) de $(E^*)^p$ dans $\mathcal{M}_p(E)$ (resp. dans S,(E), resp. dans "t,,(E)).

On dit qu'une forme p-linéaire S est un élément décomposable de $\mathcal{M}_p(E)$ (resp. de S,(E), resp. de $\mathcal{A}_p(E)$) s'il existe une suite $(y_1^*, y_2^*, ..., y_p^*)$ de p formes linéaires sur E dont le produit tensoriel (resp. symétrique, resp. extérieur) est égal à S.

Il reste à montrer que toutes les formes p-linéaires peuvent être reconstituées à l'aide des éléments décomposables lorsque l'espace vectoriel E est de dimension finie, ce qui fait l'objet du théorème fondamental suivant.

Théorème 16. Soit B = $(e_1, e_2, ... e_n)$ une base de E, et B* = $(e_1^*, e_2^*, ... e^*_n)$ la base duale de B.

1. Soit $\mathcal{F}$ l'ensemble des applications de $[1, p]$ dans $[1, n]$, et, pour tout élément $\mathcal{K}$ de $\mathcal{F}$, e_χ l'élément de $\mathcal{M}_p(E)$ défini par la formule :

$$e_\chi = e^*_{\chi(1)} \otimes e^*_{\chi(2)} \otimes ... \otimes e^*_{\chi(p)}.$$

Alors la famille $(e_\chi)_{\chi \in \mathcal{F}}$ est une basse de $\mathcal{M}_p(E)$, dite canoniquement associée à B. Par suite, la dimension de $\mathcal{M}_p(E)$ est égale à n^p, et les formes p-linéaires décomposables constituent une partie génératrice de cet espace vectoriel. En particulier, l'espace vectoriel des formes bilinéaires sur E est de dimension n^2.

2. Soit $\mathcal{S}$ l'ensemble des applications s de $[1, n]$ dans N telles que :

$$\sum_{j=1}^{n} s(j) = p,$$

et, pour tout éléments de $\mathcal{S}$, e_s l'élément de S,,(E) défini par la formule :

$$e_s = \prod_{j=1} e_j^{s(j)}.$$

Alors, si K est de caractéristique 0, la famille $(e_s)_{s \in \mathcal{S}}$ est une base de $\mathcal{S}_p(E)$, dite

canoniquement associée à B. Par suite, la dimension de S,,(E) est égale à C^p_{p+n-1}, et les formesp-linéaires symétriques décomposables constituent une partie génératrice de cet espace vectoriel. En particulier, l'espace vectoriel des formes bilinéaires symétriques sur E est de dimension $n(n+1)/2$.

3. Lorsquep > n, $\mathcal{A}_p(E) = \{0\}$. Dans le cas contraire, soit $\mathcal{T}$ l'ensemble des parties de $[1, n]$ à p éléments, et, pour tout élément P de $\mathcal{T}$, e_p l'élément de $\mathcal{A}_p(E)$ défini par la formule :

$$e_p = e_{\varphi(1)} \text{ A } e_{\varphi(2)} \text{ A } ... \text{ A } e_{\varphi(p)},$$

où φ désigne l'application strictement croissante de $[1, p]$ dans $[1, n]$ ayant P pour image. Alors la famille $(e_p)_{p\in\mathcal{T}}$ est une base de -t,(E), dite canoniquement associée à B. Par suite, la dimension de $\mathcal{A}_p(E)$ est égale à Ci;, et les formes p-linéaires alternées décomposables constituent une partie génératrice de cet espace vectoriel. En particulier, l'espace vectoriel des formes bilinéaires alternées sur E est de dimension $n(n-1)/2$.

Enfin, comme $C^n_n = 1$, l'espace vectoriel des formes n-linéaires alternées sur E est de dimension 1, résultat dont l'importance va apparaître dans la théorie des déterminants.

En utilisant la structure de $\mathcal{A}_p(E)$, on établit facilement les deux critères d'indépendance linéaire suivants, très utiles en pratique.

Soit E un espace vectoriel de dimension n sur K, etp un entier inférieur ou égal à n. Pour qu'une suite $(x_1, x_2, ... x_p)$ de vecteurs de E soit libre, il faut et il suffit qu'il existe un élément f de A,(E) tel que $f(x_1, x_2, ..., x_p) \neq 0$. De même, pour qu'une suite $(y_1^*, y_2^*, ... y_p^*)$ de formes linéaires sur E soit libre, il faut et il suffit que y_1^* A y_2^* A A $y_p^* \neq 0$.

7. Déterminants

Déterminant de n vecteurs

Soit E un espace vectoriel de dimension n sur K, et B = $(e_1, e_2, ...$ e,) une base de E. La base de $\mathcal{A}_n(E)$ canoniquement associée à B est réduite à la forme n-linéaire alternée e_1^* A e_2^* A A e_n^* ; celle-ci est la seule forme n-linéaire alternée sur E prenant la valeur 1 sur $(e_1, e_2, ...$ e,). On l'appelle *déterminant* dans la base B, et on la note $\det_B$. Pour tout élémentfde $\mathcal{A}_n(E)$ et pour toute suite $(x_1, x_2, ...$ x,) de vecteurs de E :

$$(x_1, x_2, ..., x_n) = f(e_1, e_2, ..., e_n) \det_B(x_1, x_2, \ldots, x_n).$$

Les propriétés des formes n-linéaires alternées s'appliquent à $\det_B$. De plus, le critère d'indépendance linéaire de n vecteurs s'énonce ici : pour que $(x_1, x_2, ... x_n)$ soit libre, il faut et il suffit que $\det_B(x_1, x_2, ..., x_n) \neq 0$.

Déterminant d'un endomorphisme

Soit E un espace vectoriel de dimension n sur K. Puisque $\mathcal{A}_n(E)$ est de dimension 1, tout endomorphisme de A,(E) est une homothétie. En particulier, pour tout endomorphisme U de E, l'extension U,, de U à l'espace vectoriel $\mathcal{A}_n(E)$ est une homothétie ; le rapport de cette homothétie s'appelle déterminant de l'endomorphisme U, et se note det U. Ainsi, par définition de det U, pour tout élémentfde $\mathcal{A}_n(E)$:

$$[U(x_1), U(x_2), ..., U(x_n)] = (\det U) . f(x_1, x_2, ..., x_n).$$

Des propriétés des extensions des applications linéaires on déduit les résultats suivants :

Pour tout couple (U, V) d'endomorphismes de E :

$$\det VU = (\det V) . (\det U).$$

Le déterminant de l'application identique de E est égal à 1 ; plus généralement, le déterminant de l'homothétie de rapport α est égal à α^n.
Pour qu'un endomorphisme U dc E soit inversible, il faut et il suffit que son déterminant soit non nul ; dans ces conditions :

$$\det U^{-1} = \frac{1}{\det U}.$$

Le déterminant du transposé ${}^tU$ d'un endomorphisme U de E est égal à celui de U :

$$\det {}^tU = \det U.$$

L'application $U \mapsto \det U$ est donc un morphisme du groupe linéaire GL(E) dans le groupe multiplicatif K*. Le noyau de ce morphisme est un sous-groupe distingué de GL(E), appelé *groupe spécial linéaire* de E, et noté SL(E).

Déterminant d'une matrice carrée

On appelle déterminant d'un élément A4 de M,(K), et on note det Mle déterminant de I'endomorphisme de K,, canoniquement associé à M. Les propriétés du déterminant d'un endomorphisme se transcrivent aussitôt pour les matrices. De plus, le déterminant de M n'est autre que le déterminant de ses vecteurs colonnes, ou de ses vecteurs lignes, dans la base canonique de K^n. Enfin, les matrices de déterminant 1 constituent un sous-groupe distingué de $GL_n(K)$, noté $SL_n(K)$.

En utilisant la caractérisation du déterminant dc n vecteurs, on démontre la proposition suivante : soit n et p deux entiers naturels non nuls tels que $p < n$, soit A un élément de $M_p(K)$, soit B un élément de M,, $_p$(K), soit C un élément de $M_{p,n-p}(K)$, et Ml'élément de M,,(K) défini par la formule :

$$M = \begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$$

Alors :

$$\det M = (\det A) \cdot (\det B).$$

On peut en déduire la formule de développement d'un déterminant suivant une colonne, ou une ligne. Considérons pour cela un élément $M = $ (a,,) de M,,(K), où $n > 1$; pour **tout** couple *(i, j)* d'éléments de [1, n], notons A, la matrice carrée d'ordre $n - 1$ obtenue en supprimant la i-ième ligne et la j-ième colonne de M. Alors, pour tout élément j de [1, n] :

$$\det M = \sum_{i=1}^{n} (-1)^{i+j} \alpha_{ij} \det A_{ij}.$$

On est ainsi amené à considérer la matrice M' dont les éléments α'_{ij} sont définis par la relation :

$$\alpha'_{ij} = (-1)^{i+j} \det A_{ij}.$$

La transposée de M' s'appelle matrice *complémentaire* de M, et se note $\tilde{M}$. Il en découle immédiatement que :

$$M\tilde{M} = \tilde{M} M = (\det M) I_n$$

En particulier, lorsque Mest inversible :

$$M^{-1} = \frac{1}{\det M} \tilde{M},$$

formule dont le principal intérêt est de montrer que l'application $M \mapsto M^{-1}$ est rationnelle.

8. Produits tensoriels

Produit tensoriel d'espaces vectoriels

La notion de produit tensoriel sert a remplacer l'étude des applications multinéaires par celle des applications linéaires. Plus précisément, on obtient le résultat suivant.

Théorème 17. Soit $E_1, E_2, \ldots . E_p$ des espaces vectoriels sur K. Il existe un couple (G, T) constitué d'un espace vectoriel G sur K et d'une application multilinéaire T de E, X E_2 X ,,, X E_p dans Ci possédant la propriété universelle suivante : Pour tout couple (F, S) constitué d'un espace vectoriel F sur K et d'une application multilinéaire S de E, X E_2 X X E_p dans F, il existe une application linéaire $\tilde{S}$ et une seule de G dans F telle que S = $\tilde{S} \circ$ T. Un tel couple (G,T) est unique à isomorphisme près. L'espace vectoriel G s'appelleproduit *tensoriel* des espaces vectoriels $E_1, E_2, \ldots .$ E_p, et se note $E_1 \otimes E_2 \otimes \otimes E_p$. L'application multilinéaire T se note :

$$(x_1, x_2, \ldots, x_p) \mapsto x_1 \otimes x_2 \otimes \otimes x_p.$$

L'application S $\mapsto$ S est un isomorphisme de l'espace vectoriel $\mathcal{M}(E_1$ X E_2 X ... X E_p, F) des applications multilinéaires de E, X E_2 X X E_p dans F sur l'espace vectoriel C(E, $\otimes$ $E_2 \otimes \otimes E_p$, F). Les éléments de E, $\otimes$ $E_2 \otimes \otimes E_p$ de la forme $x_1 \otimes x_2 \otimes \otimes x_p$ sont dits décomposables ; ils constituent une partie génératrice de E, $\otimes E_2 \otimes \ldots \otimes E_p$. Si, pour tout élémentjde [1 ,p], E, est de dimension finie n_j et est muni d'une base B, = (e_{ij}), alors les éléments $e_{i_1,1} \otimes e_{i_2,2} \otimes \ldots \otimes e_{i_p,p}$ constituent une base de E, $\otimes$ $E_2 \otimes \otimes E_p$, dite canoniquement associée aux bases B,. En particulier :

$$\dim E_1 \otimes E_2 \otimes \ldots \otimes E_p = n_1 n_2 \ldots n_p.$$

Soit maintenant (E,, $E_2, \ldots .$ E,,) et (F_1, $F_2, \ldots .$ F,,) deux suites d'espaces vectoriels sur K, et, pour tout élément j de [1, p], U, un élément de C(E,, F,). Il existe alors une application linéaire U et une seule de $E_1 \otimes E_2 \otimes \ldots \otimes E_p$ dans F, $\otimes$ $F_2 \otimes \otimes F_p$ telle que, pour tout élément (x_1, x_2, x_p) de $E_1 \otimes E_2 \otimes \ldots \otimes$ E,, :

$$U(x_1 \otimes x_2 \otimes \ldots \otimes x_p) = U_1(x_1) \otimes U_2(x_2) \otimes \otimes U_p(x_p).$$

On l'appelle produit tensoriel des applications linéaires U,, et on la note U, $\otimes$ U_2 $\otimes$ $\otimes$ u,.

Trace d'un endomorphisme

Soit E et F deux espaces vectoriels sur K. Pour tout élément (a*, b) de E* X F, l'application $U_{a^*,b}$ qui à tout vecteur x de E associe le vecteur < a*, x > b de F est une application linéaire de E dans F ; on l'appelle application linéaire élémentaire associée à (a^*, h). Si a^* et b ne sont pas nuls, l'image de $U_{a^*,b}$ est la droite Kb de F, et son noyau est l'hyperplan de E noyau de la forme linéaire a*. De plus, l'application $(a^*, b) \mapsto U_{a^*,b}$ est une application bilinéaire de E* X F dans C(E, F).

Il existe une application linéairej et une seule de E* $\otimes$ F dans C(E, F) telle que, pour tout élément (y^*, z) de E* X F :

$$j(y^* \otimes z) = U_{y^*,z}.$$

En effet, l'application $(y^*, z) \mapsto U_{y^*,z}$ est une application bilinéaire de E* X F dans C(E, F). La propriété universelle du produit tensoriel E* $\otimes$ F montre alors l'existence et l'unicité de j.

De plus, si E et F sont de dimension finie, j est un isomorphisme, car j est injective et que :

$$\dim E^* \otimes F = (\dim E^*) \cdot (\dim F).$$

Soit enfin E un espace vectoriel sur K. 11 existe une forme linéaire c et une seule sur l'espace vectoriel E* $\otimes$ E telle que, pour tout élément (y^*, x) de E* X E :

$$c(y^* \otimes x) = \langle y^*, x \rangle.$$

L'application c s'appelle contraction canonique de E* $\otimes$ E dans K.

En effet, l'application $(y^*, x) \mapsto$ < y^*, x > est une application bilinéaire de E* X E dans K. La propriété universelle

du produit tensoriel $E^* \otimes E$ prouve l'existence et l'unicité de c.

En combinant les résultats précédents, nous obtenons le théorème qui suit.

Théorème 18. Soit E un espace vectoriel de dimension finie sur K. Il existe une forme linéaire et une seule sur l'espace vectoriel L(E), appelée *trace* et notée tr, telle que, pour tout endomorphisme élémentaire $U_{a^*,b}$ de E,

$$\operatorname{tr} U_{a^*,b} = \langle a^*, b \rangle.$$

Cette forme linéaire n'est autre que $c \circ j^{-1}$, où j désigne l'isomorphisme canonique de $E^* \otimes E$ sur $\mathcal{L}(E)$.

De plus, la trace possède les propriétés suivantes :

– Pour tout couple (U, V) d'endomorphismes de E :

$$\operatorname{tr}(VU) = \operatorname{tr}(UV).$$

Pour tout endomorphisme U de E :

$$\operatorname{tr} {}^tU = \operatorname{tr} U.$$

Enfin, on définit la trace d'un élément $M = (\alpha_{ij})$ de $M_n(K)$ comme la trace de l'endomorphisme de K^n canoniquement associé à M ; on vérifie que :

$$\operatorname{tr}(M) = \sum_{i=1}^{n} \alpha_{ij}.$$

9. Modules

Soit A un anneau unitaire. On appelle A-*module* à gauche un ensemble E muni de deux lois de composition satisfaisant aux mêmes axiomes que les espaces vectoriels. On définit de même les A-modules à droite : cette fois

$$\alpha . (\beta . x) = (\beta\alpha) . x.$$

Par exemple, l'application $(n, x) \mapsto nx$ définit sur tout groupe abélien une structure de Z-module.

Les résultats des chapitres 1 et 2 s'étendent sans changement dans ce cadre plus général, à ceci près que, lorsque l'anneau A n'est pas commutatif, les homothéties ne sont pas des endomorphismes, si bien qu'il n'est plus possible de munir le groupe additif $\mathcal{L}(E, F)$ d'une structure de A-module et l'anneau C(E) d'une structure de A-algèbre. Enfin, le dual d'un A-module à gauche doit être considéré comme un A-module à droite.

Existence de bases

Une différence essentielle avec les espaces vectoriels est la suivante : il peut arriver qu'une partie réduite à un élément non nul ne soit pas libre. C'est le cas pour les éléments de $\mathbf{Z}/n\mathbf{Z}$, considéré comme Z-module.

De plus, alors que, dans tout espace vectoriel, il existe des bases (cf. théorème 8), il n'en est pas de même dans tout module, même lorsqu'il existe une partie génératrice réduite à un seul élément ; c'est le cas pour $\mathbf{Z}/n\mathbf{Z}$. Un module admettant une base est dit *libre.*

Existence de supplémentaires

De même, le théorème 9 ne se généralise pas à tous les modules. Ainsi, le sous-module du Z-module Z engendré par 2 n'admet pas de sous-module supplémentaire. Un sous-module admettant un supplémentaire est appelé *facteur direct.*

On dit qu'un A-module E est *semi-simple* si **tout** sous-module de E est un facteur direct. La théorie des modules semi-simples est utile pour la réduction des endomorphismes : Soit en effet U un endomorphisme d'un espace vectoriel E, et A le sous-anneau de L(E) engendré par U. L'application $(V, x) \mapsto V(x)$ fait de E un A-module. Les sous-modules de E ne sont autres que les sous-espaces vectoriels

de E stables par U. Pour que le module E soit semi-simple, il faut et il suffit que tout sous-espace vectoriel stable par U admette un supplémentaire stable. On dit alors que U est semi-simple.

Plus généralement, la théorie des modules semi-simples est utile pour la représentation linéaire des groupes (cf. GROUPES - Groupes de Lie), où elle intervient sous le nom de complète réductibilité.

Modules de type fini

On dit qu'un A-module E est de type fini s'il existe une partie génératrice finie de E. (Ici, la terminologie « de dimension finie » serait désastreuse, puisqu'un module de type fini peut très bien ne pas avoir de base.) Même lorsque E est un A-module libre de type fini, il peut arriver qu'il existe deux bases finies de E n'ayant pas le même nombre d'éléments. Cependant, ce phénomène ne se produit pas lorsque l'anneau A est commutatif.

On peut envisager deux généralisations « raisonnables » du théorème 11 concernant les sous-espaces vectoriels des espaces vectoriels de dimension finie :

1. Un sous-module d'un A-module de type fini n'est pas, en général, de type fini ; c'est cependant le cas lorsque l'anneau A est noethérien. L'importance de ce cas apparaît dans la théorie des polynômes (cf. ANNEAUX COMMUTATIFS) et en géométrie algébrique.

2. Un sous-module d'un A-module libre n'est pas nécessairement libre ; c'est cependant le cas lorsque l'anneau A est principal. Les résultats sont riches en applications pour la théorie des groupes abéliens et la réduction des endomorphismes, à propos des diviseurs élémentaires (cf. théorie SPECTRALE).

Applications multilinéaires et déterminants

Les résultats des chapitres 6 et 7 s'étendent sans changement au cas des A-modules libres de type fini sur un anneau commutatif, sauf les critères d'indépendance linéaire et d'inversibilité. En particulier, pour qu'un endomorphisme U soit inversible, il faut et il suffit que det U soit inversible dans l'anneau A. Ainsi modifié, ce critère d'inversibilité permet d'étudier les équations linéaires à coefficients entiers et, plus généralement, suivant les méthodes de Dedekind et de Kronecker, les entiers algébriques.

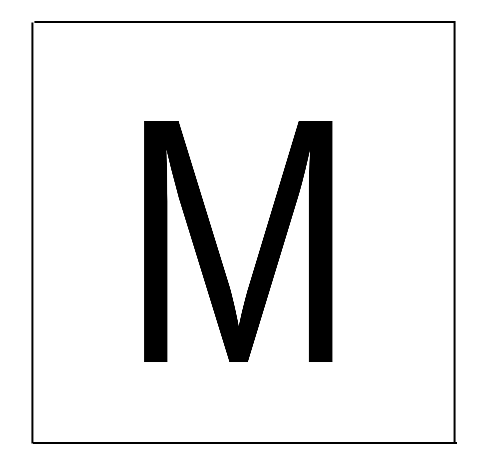
LUCIEN CHAMBADAL et JEAN-LOUIS OVAERT

Bibliographie

N. BOURBAKI, *Algèbre* : chap. I à III, Masson, Paris, 1982 / R. CABANE & C. LEBŒUF, *Algèbre linéaire*, 2 vol., Ellipses, Paris, 1987-1990 / L. CHAMBADAL & J.-L. OVAERT, *Algèbre multilinéaire*, Dunod, Paris, nouv. éd. 1984 / R. GANTMACHER, *Théorie des matrices*, 2 vol.. *ibid.*, 1966, reprod. fac-sim.. Gabay, Sceaux, 1990 / R. GODEMENT. *Cours d'algèbre*, Hermann, Paris, 1963 / W. GRAEUB. *Linear Algebra*, Springer, Berlin, 4' éd. 1991 / J. GRIFONE, *Algèbre linéaire*, Cepadues, Toulouse, 1990 / N. JACOBSON, Basic *Algebra*, 2 vol., W. H. Freeman, New York, 2' éd. 1984-1985 / S. LANG. *Algebra*, Springer-Verlag, New York, 3' éd. 1989 / J.-L. OVAERT & J.-L. VERLEY. *Algèbre*, vol. 1, C.E.D.I.C.-Nathan, Paris, 198 1 / G. W. STEWART, *Introduction to Matrix Computations*, Acad. Press, New York, 1973 / B. L. VAN DER WAERDEN, *Modern Algebra*, 2 vol., Springer-Verlag. New York, 1990.

LOGARITHME

→ EXPONENTIELLE & LOGARITHME



MESURE → INTÉGRATION & MESURE

MÉTRIQUES ESPACES

La notion d'espace métrique, introduite en 1906 par M. Fréchet et développée peu après par F. Hausdorff, est directement issue d'une analyse des principales propriétés de la distance usuelle. L'extension aux espaces métriques des propriétés de l'espace euclidien qui sont définissables à partir de la distance seule introduit un langage géométrique dans de nombreuses questions d'analyse et de théorie des nombres. C'est ainsi que l'on définit, à partir des boules, les ouverts. Par la manière naturelle dont s'introduisent les voisinages et les notions de limite et de continuité, l'étude des espaces métriques est une excellente introduction à la topologie générale.



1. Distances

L'analyse des principales propriétés de la distance entre deux points dans l'espace euclidien conduit à la définition axiomatique suivante. On appelle *distance* sur un ensemble E une application d de E X E dans l'ensemble $\mathbf{R}_+$ des nombres réels positifs ou nul telle que, quels que soient les éléments x, y et z de E, on ait :

(1) $d(x,y) = 0 \Leftrightarrow x = y$,

(2) $d(x,y) = d(y,x)$,

(3) $d(x,z) a d(x,y) + d(y,z)$;

cette dernière condition est appelée *inégalité trianguhire* car elle est la généralisation de la classique inégalité entre les longueurs des côtés d'un triangle.

Un ensemble E muni d'une distance s'appelle un *espace métrique*. Si (E, d) et (E', d') sont deux espaces métriques, une bijection f de E sur E' sera dite une *isométrie* si elle conserve la distance, c'est-à-dire si $d'(f(x), f(y)) = d(x, y)$ quels que soient $x, y \in$ E ; deux espaces métriques sont dits *isométriques* s'il existe une telle isométrie de l'un sur l'autre et présentent alors, « par transport » au moyen de cette isométrie, des propriétés semblables.

Exemples

On verra dans ce qui suit que la notion d'espace métrique recouvre un matériau mathématique très varié. Comme exemple extrême, remarquons que tout ensemble peut être muni de la distance, dite *triviale,* définie par $d(x, x) = 0$, $d(x, y) = 1$ si $x \neq y$. Si E est un espace métrique de distance d, tout sous-ensemble A de E est un espace métrique, dit *sous-espace métrique* de E

pour la distance induite d' définie par $d'(x, y) = d(x, y)$, $x, y \in A$.

Une classe très importante d'espaces métriques est constituée par les *espaces vectoriels normés,* en définissant ici la distance de deux éléments x et y comme la norme de leur différence, soit :

$$d(x,y) = \|x - y\|;$$

la distance ainsi obtenue est invariante pour les translations de l'espace vectoriel, c'est-à-dire $d(x - a, y - a) = d(x, y)$ quels que soient les éléments x, y et a. Nous renvoyons à l'article espaces vectoriels NORMÉS pour de nombreux exemples, relatifs à l'analyse fonctionnelle notamment, et mentionnerons seulement ici les trois distances suivantes, qui sont déduites des normes correspondantes, sur $\mathbf{R}^2$:

$$d_1(x,y) = |x_1 - y_1| + |x_2 - y_2| = \|x - y\|_1,$$
$$d_2(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} = \|x - y\|_2,$$
$$d_3(x,y) = \sup(|x_1 - y_1|, |x_2 - y_2|),$$

où $x = (x_1, x_2)$, $y = (y_1, y_2)$. Ces distances vérifient les inégalités :

$$d_3 \leqslant d_2 \leqslant d_1 \leqslant 2 d_3,$$

dont on verra des conséquences plus bas. Si (E. 6) et (E', 6') sont des espaces métriques, on peut utiliser ce qui précède pour définir des distances 6, sur le produit cartésien E X E', en posant :

$$\delta_i((x, x'),(y, y')) = d_i(\delta(x,y), \delta'(x', y')),$$

où $i = 1, 2, 3$, et où les d_i sont les distances sur $\mathbf{R}^2$ ci-dessus.

Les espaces vectoriels normés sont les espaces métriques dont les propriétés « ressemblent le plus » à celles des espaces numériques habituels. Donnons maintenant des exemples qui ne rentrent pas dans ce cas.

Lu droite numérique achevée. Désignons par $\overline{\mathbf{R}}$ la droite numérique achevée, $\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty\} \cup \{+\infty\}$, qui est obtenue en adjoignant à l'ensemble R des nombres réels deux nouveaux éléments, que l'on désigne traditionnellement par $-\infty$ et $+\infty$ vu le rôle qu'ils jouent en analyse, et remarquons que l'application f, définie par :

$$f(x) = \frac{x}{1 + |x|}, \quad x \in \mathbf{R},$$
$$f(+\infty) = +1, \quad f(-\infty) = -1,$$

réalise une bijection de $\overline{\mathbf{R}}$ sur le segment $[-1, +1]$. On peut donc transporter la distance usuelle sur R, en définissant une distance d sur $\overline{\mathbf{R}}$ par :

$$d(x,y) = |f(x) - f(y)|;$$

bien entendu f est une isométrie de $\overline{\mathbf{R}}$, muni de cette distance d, sur le segment fermé $[-1, 1]$ muni de la distance habituelle, puisqu'on a fait exactement ce qu'il fallait pour cela !

– *Distances p-udiques sur* Q. En théorie des nombres, on associe à tout nombre premier p une distance sur l'ensemble Q des nombres rationnels de la manière suivante. Pour tout entier n strictement positif, soit $v(n)$ sa valuation p-adique, c'est-à-dire l'exposant de p dans la décomposition de n en facteurs premiers ; ainsi $u(n) = 0$ pour n non divisible par p et $v(nn') = v(n) + v(n')$. Prolongeons alors v à l'ensemble Q* des rationnels non nuls en remarquant que si $x = \pm r/s$, alors $v(x) = u(r) - v(s)$ ne dépend que du rationnel x et non du choix de la fraction r/s ; on vérifie facilement que l'on a :

(a) $v(-z) = v(z)$,
(b) $v(z + z') \leqslant \inf(v(z), v(z'))$,

pour z, $z' \in \mathbf{Q}^*$. On définit alors la distance p-adique sur Q par :

$$d(x,x) = 0,$$
$$d(x,y) = p^{-v(x-y)},$$

où x, $y \in \mathbf{Q}$. La condition (1) est claire, et la condition (2) résulte de (a). La condition (b) entraîne que l'on a la condition (3′), qui entraîne (3) :

(3′) $\quad d(x,z)$ 4 $\sup(d(x,y), d(y,z))$.

Un espace métrique dont la distance vérifie la condition (3′), plus forte que l'inégalité triangulaire (3), est dit *ultramétrique* ; comme on le verra, ces espaces ont des propriétés très particulières.

Le langage géométrique

Les boules, définies à partir de la distance comme dans l'espace euclidien, constituent la notion géométrique essentielle dans les espaces métriques. Dans un espace métrique E de distance d, on appelle *boule ouverte* de centre $x_0 \in$ E et de rayon $r > 0$, l'ensemble des points de E dont la distance à x_0 est strictement inférieure à **r**, soit :

$$B(x_0, r) = \{x \in E \mid d(x_0, x) < r\} ;$$

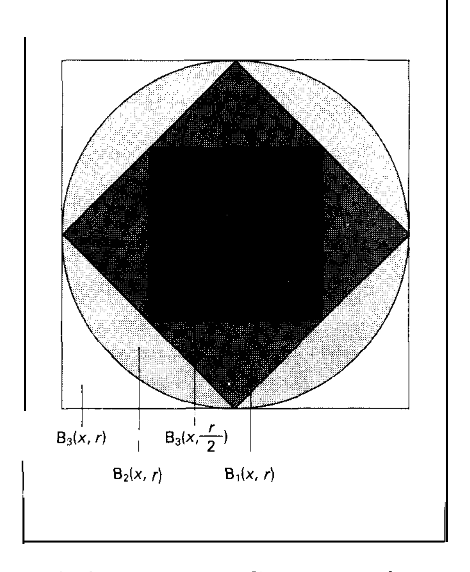
de manière analogue, on définit la *boule fermée* de centre x_0 et de rayon r :

$$B_f(x_0, r) = \{x \in E \mid d(x_0, x) \leqslant r\} ;$$

la justification des qualificatifs ouvert et fermé apparaîtra plus bas. Nous renvoyons aux figures 1, 2 et 3 de l'article sur les espaces vectoriels NORMÉS pour une représentation des boules quand on munit $\mathbf{R}^2$ des distances d_1, d_2 et d_3 déjà mentionnées. Les inégalités établies entre ces distances se traduisent par les inclusions :

$$B_3\left(x, \frac{r}{2}\right) \subset B_1(x,r) \subset B_2(x,r) \subset B_3(x,r),$$

en désignant par B_i la boule pour la distance d_i (cf. figure).



Si A est une partie d'un espace métrique E et x $\in$ E, on appelle *distance* de x à A la borne inférieure de l'ensemble des nombres $d(x, y)$ pour y $\in$ A, soit :

$$d(x, A) = \inf_{y \in A} d(x,y) ;$$

c'est le rayon de la plus grande boule ouverte de centre x qui ne rencontre pas A. Pour bien expliquer cette notion, établissons la relation :

$$|d(x,A) - d(x',A)| \; 6 \; d(x,x'),$$

pour deux éléments x et x' quelconques de E. Par définition de la borne inférieure, pour tout $\varepsilon > 0$, il existe $y \in$ A tel que :

$$d(x,y) \leqslant d(x,A) + \varepsilon ;$$

par suite, d'après l'inégalité triangulaire,

$$d(x',y) \leqslant d(x',x) + d(x,y)$$
$$\leqslant d(x,x') + d(x,A) + \varepsilon,$$

ce qui entraîne :

$$d(x',A) \leqslant d(x',y) \leqslant d(x,x') + d(x,A) + \varepsilon,$$

donc :

$$d(x',A) \leqslant d(x,x') + d(x,A) + \varepsilon,$$

et, par suite, puisque ε est quelconque,

$$d(x',A) - d(x,A) \leqslant d(x,x');$$

échangeant les rôles de x et x' et rassemblant les deux résultats, on obtient la conclusion recherchée.

Continuons à transposer le vocabulaire géométrique usuel ; nous dirons qu'un sous-ensemble A d'un espace métrique E de distance d est *borné* s'il est contenu dans une boule. Cela équivaut à dire que le nombre :

$$d(A) = \sup_{x,y \in A} d(x,y),$$

appelé *diamètre* de l'ensemble A, est fini. Le langage géométrique que nous venons d'introduire dans les espaces métriques constitue un support pour une intuition géométrique qui est très utile mais doit être soigneusement contrôlée car elle n'est pas sans surprises et risque d'utiliser implicitement, sans s'en rendre compte, des propriétés de l'espace euclidien plus riches que sa seule structure métrique. Par exemple, si A est une boule de rayon r, on montre facilement avec l'inégalité triangulaire que son diamètre est $\leqslant 2r$, mais on ne peut affirmer l'égalité que si on a un espace vectoriel normé. Les espaces ultramétriques présentent ainsi de nombreux phénomènes « pathologiques », en ce sens qu'ils sont contraires à l'intuition géométrique courante ; le mathématicien qui travaille sur ces espaces est ainsi amené à développer une « intuition géométrique » qui peut sembler tout à fait ésotérique au profane... Ainsi, dans un tel espace, il n'y a pas de boules *sécantes*, en ce sens que si deux boules $B(x, r)$ et $B(x', r')$, $r \leqslant r'$, ont un point commun y, alors $B(x, r) \subset B(x', r')$. En effet, soit z un point quelconque de $B(x, r)$, c'est-à-dire $d(x, z) < r \leqslant r'$; on a, d'après (3′), $d(y, z) \leqslant \sup(d(y, x), d(x, z)) < r$ puisqu'on a aussi $d(x, y) < r$ car $y \in B(x, r)$. Utilisant à nouveau (3′), on a :

$$d(x', z) \leqslant \sup(d(x', y), d(y, z)) < r'$$

donc $z \in B(x', r')$, ce qui établit l'inclusion annoncée ; on aurait une démonstration analogue pour deux boules fermées. Nous verrons dans le chapitre suivant que cette particularité des espaces ultramétriques nous réserve encore bien des surprises.

2. Topologie d'un espace métrique

À partir des boules, on peut construire sur un espace métrique les principales notions topologiques qui permettent de « faire de l'analyse ». À ce propos, par la clarté avec laquelle les notions de limite et de continuité s'expriment au moyen de la terminologie que nous allons introduire, la théorie des espaces métriques constitue un excellent préliminaire à la topologie générale.

Ouverts et fermés

Soit E un espace métrique de distance d. On dit qu'un sous-ensemble U de E est *ouvert* si pour tout point $x \in$ U il existe une boule ouverte de centre x contenue dans U. D'après un principe général de logique, l'ensemble vide, qui n'a pas d'élément, est donc ouvert. Faisons le lien avec la terminologie introduite plus haut en montrant qu'*une boule ouverte* $B(x_0, r)$ *est un ensemble ouvert* : en effet, si $x \in B(x_0, r)$, l'inégalité triangulaire entraîne que $B(x, r')$

C $B(x_0, r)$ pour $r' = r - d(x_0, x) > 0$. On voit donc qu'un ensemble U est ouvert si et seulement si c'est une réunion de boules ouvertes.

La famille des ouverts d'un espace métrique vérifie les propriétés suivantes qui sont prises en topologie générale comme axiomes pour définir une topologie : (0,) E et ϕ sont des ensembles ouverts ; (O_2) Toute réunion (finie ou pas) d'ensembles ouverts est un ensemble ouvert; (0,) Toute intersection *finie* d'ouverts est un ouvert.

Montrons ce dernier point. Soit U,, U_2, ... U_n des ouverts ; si l'intersection est vide, c'est terminé d'après (0,). Sinon, soit $x \in$ u , $\cap U_n$ = U ; par hypothèse, il existe $r,, i = 1, \ldots n$, tels que $B(x, r_i)$ C U_i et par suite $B(x, r)$ C U pour $r = \inf (r_1, \ldots, r_n)$.

On dit que deux distances d_1 et d_2 sur un même ensemble E sont (topologiquement) équivalentes si les ouverts correspondants sont les mêmes. Cela signifie que toute boule ouverte B_1 de centre x_0 par rapport à la distance d_1 contient une boule B_2 de centre x_0 (par rapport à la distance d_2), puisque $x_0 \in$ B,, qui, étant un ouvert pour d_1, est aussi un ouvert pour d_2, et vice versa. C'est ce qui se produit pour les trois distances d,, d_2 et d_3 considérées plus haut sur $\mathbf{R}^2$, qui définissent les mêmes ouverts (cf. figure). Remarquons aussi que, sur R, la distance usuelle d définie à partir de la valeur absolue par $d(x, y) = |x - y|$, et pour laquelle les boules sont les intervalles, est topologiquement équivalente à la distance d' pour laquelle R est un sous-espace métrique de $\overline{\mathbf{R}}$ (cf. chap. 1), c'est-à-dire :

$$d'(x, y) = \left| \frac{x}{1 + |x|} - \frac{y}{1 + |y|} \right|$$

Par passage au complémentaire, on définit la famille des fermés d'un espace métrique : un sous-ensemble F de E est dit *fermé* si son complémentaire dans E est un ensemble ouvert. Par exemple *toute boule fermée* $B_f(x_0, r)$ *est un ensemble fermé* ; en effet, si $x \notin B_f(x_0, r)$, on a :

$$B(x, r') \cap B_f(x_0, r) = 0$$

pour $r' = d(x_0, x) - r > 0$, comme cela résulte facilement de l'inégalité triangulaire. Des propriétés (O,), (0,) et (0,) il résulte facilement que ϕ et E sont des fermés ; toute intersection (finie ou pas) de fermés est un fermé ; toute réunion *finie* de fermés est un fermé. Si A est un sous-ensemble quelconque de E, il existe donc un « plus petit » (pour l'inclusion) ensemble fermé contenant A, à savoir l'intersection $\overline{A}$ de la famille de tous les fermés qui contiennent A ; cet ensemble $\overline{A}$ est aussi le complémentaire du « plus grand » (toujours pour l'inclusion) ensemble ouvert ne rencontrant pas A, qui est la réunion des boules ouvertes ne rencontrant pas A. Ainsi, un point $x \in$ E appartient à $\overline{A}$ si et seulement si toute boule ouverte de centre x rencontre A, ce qui revient au fait que la distance de x à A est nulle : on dit alors que x est un *point adhérent* à A. L'ensemble fermé $\overline{A}$ s'appelle la *fermeture, ou l'adhérence* de A. On dit que A est partout dense dans E si $\overline{A}$ = E ; par exemple l'ensemble Q des nombres rationnels est partout dense dans R.

Les notions précédentes sont la transposition, au moyen des boules, de notions familières dans les espaces numériques et, comme telles, sont aussi l'objet d'un investissement intuitif qui n'est cependant pas toujours justifié. Ainsi on pourrait penser que l'adhérence d'une boule ouverte est toujours la boule fermée de même centre et de même rayon mais, si cela est vrai pour les espaces vectoriels normés, les espaces ultramétriques nous réservent ici encore

des surprises. En effet, montrons que, dans un tel espace, toute boule ouverte $B(x_0, r)$, qui est un ensemble ouvert comme nous l'avons vu, est aussi un ensemble fermé et, par suite, est sa propre adhérence. Il suffit pour cela de remarquer que si $x \notin B(x_0, r)$, alors la boule $B(x, r/2)$ ne rencontre pas la boule $B(x_0, r)$ car sinon, puisqu'il n'y a pas de boules « sécantes ». la boule de rayon $r/2$ serait contenue dans l'autre (cf. fin du chap. 1), d'où, en particulier, $x \in B(x_0, r)$, ce qui contredit l'hypothèse sur x.

Voisinages et continuité

Introduisons maintenant les voisinages pour préciser la notion de continuité. Soit E un espace métrique de distance d. On dit qu'un ensemble V C E est un *voisinage* d'un point $x \in$ E s'il contient un ouvert contenant x ; cette notion donc est « topologique » : elle ne dépend que des ouverts de l'espace métrique E, ouverts caractérisés, à leur tour, par le fait qu'ils sont voisinages de chacun de leurs points. En particulier, les boules ouvertes de centre x sont des voisinages de x et il en est de même des boules fermées puisqu'on a toujours $B(x, r)$ C $B_f(x, r)$. Les boules ouvertes (ou les boules fermées) de centre x constituent *un système fondamental de voisinages* de x, en ce sens qu'un ensemble est un voisinage de x si et seulement s'il contient une telle boule ; on obtient un système fondamental *dénombrable* de voisinages en se limitant par exemple aux boules **de** rayon $1/n$, avec n entier positif.

Soit maintenant E et E' deux espaces métriques de distances respectives d et d' etfune application de E dans E'. Analysons la notion de continuité telle qu'elle est suggérée dc manière naturelle par la définition classique (quand E = E' = R muni de la distance usuelle définie à partir de la valeur absolue) : on dit quef'est *continue* en *un point* $x_0 \in$ E si, pour tout nombre réel $\varepsilon > 0$, il existe un nombre réel $\eta > 0$ tel que :

$$d(x_0, x) < \eta \Rightarrow d'(f(x_0), f(x)) < \varepsilon,$$

ce qui exprime l'inclusion :

$$B(x_0, \eta) \subset f^{-1}(B'(f(x_0), \varepsilon)) ;$$

or, dire qu'il existe η tel que l'on ait cette inclusion signifie que l'image réciproque par f de la boule $B'(f(x_0), \varepsilon)$ est un voisinage de x_0 dans E. Puisqu'une partie de E' est un voisinage de $f(x_0)$ si et seulement si elle contient une boule de centre $f(x_0)$, on peut maintenant donner, en termes de voisinages seuls, la définition suivante de la continuité *: f est continue en x_0 si et seulement si l'image réciproque pur f de tout voisinage de f (x_0) est un voisinage de x_0.*

Il en résulte que f est continue en tout point de E (on dit que f est continue, sans préciser davantage) si et seulement si l'image réciproque par f de tout ouvert de E' est un ouvert de E.

Les notions précédentes sont topologiques : elles ne dépendent que des ouverts des espaces considérés. 11 est utile d'avoir une notion de continuité plus forte, qui n'est plus seulement topologique, utilisant le fait que l'on peut « comparer » les voisinages de deux points distincts grâce aux rayons des boules. Plus précisément, on dira qu'une application f: E → E' est *uniformément continue* si, pour tout $\varepsilon > 0$. il existe $\eta > 0$ tel **que :**

$$d(x,y) < \eta \Rightarrow d'(f(x),f(y)) < \varepsilon,$$

ou encore :

$$B(x, \eta) \subset f^{-1}(B'(f(x), \varepsilon)),$$

pour touts $\in$ E. Un cas particulier de cette situation est fourni par les applications

lipschitziennes : on dit que f est *lipschitzienne* de rapport k si on a :

$$d'(f(x),f(y)) \leqslant k\, d(x,y),$$

quels que soient x, y dans E ; ainsi l'image réciproque parfde la boule de centre $f(x)$ et de rayon r contient la boule de centre x et de rayon r/k. Par exemple, on a vu plus haut que, pour toute partie A de E, l'application qui à x $\in$ E associe sa distance à A est lipschitzienne de rapport 1.

Le langage des suites

Soit (u,) une suite de points d'un espace métrique E (de distance d). On dira de manière naturelle que cette suite converge vers un élément a $\in$ E pour n tendant vers l'infini si $d(a,u_n) \longrightarrow 0$ pour $n \longrightarrow \infty$. Pour tout ε > 0, il existe donc un entier N tel que :

$$n \geqslant \mathbf{N} \Rightarrow d(a,u_n) < \varepsilon,$$

c'est-à-dire :

$$n \geqslant \mathbf{N} \Rightarrow u_n \in \mathbf{B}(a,\varepsilon)\,;$$

ainsi $u_n \longrightarrow$ a pour $n \longrightarrow \infty$ si, pour tout voisinage V de a, il existe un entier N tel que $u_n \in$ V pour $n \geqslant$ N. On peut dire aussi que V contient les u_n sauf pour un nombre *fini* d'entiers n. Remarquons que la notion de suite convergente est topologique.

La définition précédente généralise bien entendu la notion de limite d'une suite dans l'espace $\mathbf{R}^n$ pour l'une quelconque des distances équivalentes habituelles. À titre d'exemple, cherchons maintenant à quelle condition une suite (u,,) de nombres réels positifs va converger vers l'élément + ∞ dans la droite réelle achevée $\overline{\mathbf{R}}$ munie de la distance d décrite dans le chapitre 1. On a :

$$d(+\infty,x) = 1 - \frac{x}{1+x} = \frac{1}{1+x},$$

pour tout $x \geqslant 0$. Soit $\varepsilon > 0$; on aura $d(+\infty, u_n) < \varepsilon$, pour $u_n > (1/\varepsilon) - 1$. Ainsi $u_n \longrightarrow +\infty$ pour $n \longrightarrow \infty$ si et seulement si, pour tout M = $(1/\varepsilon) - 1$, il existe un entier N tel que :

$$n \geqslant \mathbf{N} \Rightarrow u_n > \mathbf{M}\,;$$

on retrouve la notion classique de suite de nombres réels « tendant vers l'infini » pour $n \longrightarrow \infty$.

Dans un espace métrique E de distance d, le fait que tout point possède un système fondamental *dénombrable* de voisinages permet d'exprimer toutes les propriétés topologiques en termes de suites. Montrons par exemple qu'un *point x_0 est adhérent à une purtie* A *de* E *si et seulement s'il est limite dans* E *d'une suite de points de* A : pour x adhérent à A, si on choisit pour tout entier positif n un point u_n dans A $\cap$ $\mathbf{B}(x_0, 1/n)$, on obtient une suite de points de A qui converge vers x_0, puisque $d(x_0,u_n) < 1/n$; réciproquement, si (u_n) est une suite de points de A qui converge vers x_0, tout voisinage V de x_0 contient tous les termes de la suite pour n assez grand et par suite A $\cap$ V $\neq \emptyset$. Un ensemble A est donc fermé si et seulement s'il contient les limites de toutes ses suites qui sont convergentes dans E ; ainsi, on peut définir les fermés, et par complémentarité les ouverts, à partir de la notion de suite convergente. De même, on montre qu'on peut caractériser la continuité d'une applicationfen termes de suites : *une application f d'un espace métrique dans un autre est continue en un point x_0 si et seulement si, pour toute suite (u,) qui converge vers x_0, la suite image $(f(u_n))$ converge vers $f(x_0)$*. La compacité va nous donner d'autres exemples de l'importance de la notion de suite.

Espaces métriques compacts

On montre que tout sous-ensemble fermé et borné C de l'espace numérique $\mathbf{R}^n$ possède la propriété suivante, appelée

propriété de Borel-Lebesgue : pour toute famille (U,) d'ouverts de $\mathbf{R}^n$ dont la réunion contient C (on dit qu'on a un recouvrement ouvert de C), il existe une sous-famille *finie* $U_{i_1}, \dots, U_{i_n}$ dont la réunion contient C. L'importance de cette propriété dans de nombreuses questions fines d'analyse mathématique a conduit à étudier systématiquement les espaces topologiques qui la possèdent (cf. TOPOLOGIE GÉNÉRALE).

On dit qu'un espace métrique E est *compact* si, pour toute famille d'ouverts de E dont la réunion est égale à E, il existe une sous-famille finie possédant cette propriété. Ainsi, tout fermé borné C de $\mathbf{R}^n$, muni de la distance induite par celle de $\mathbf{R}^n$, est un espace métrique compact (car les ouverts du sous-espace métrique C sont les intersections avec C des ouverts de $\mathbf{R}^n$) ; en particulier, pour $n = 1$, le segment fermé $[-1, +1]$ est compact et il en est donc de même de la droite numérique achevée $\overline{\mathbf{R}}$ qui est isométrique à ce segment (cf. chap. 1).

Tout espace métrique compact est de diamètre fini, ce qui exprime que la distance entre deux points d'un tel espace est bornée ; on montre que tout sous-espace métrique fermé d'un espace métrique *compact* est compact.

Avant de donner, en termes de suites, une caractérisation des espaces métriques compacts, il nous faut introduire une nouvelle notion. On dit qu'un point a est un *point d'accumulation* d'une suite (u_n) si tout voisinage de a contient les points u_n pour une infinité d'entiers n ; ainsi, si la suite (u,,) converge vers a, ce point a est le seul point d'accumulation, mais il peut exister plusieurs points tels, chacun d'entre eux étant caractérise par le fait qu'on peut trouver une *sous-suite* de la suite initiale qui converge vers ce point.

Nous pouvons maintenant énoncer la *propriété de Bolzano- Weierstrass*, vraie pour tout espace topologique compact, mais qui *caractérise* les espaces métriques *compacts* parmi les espaces métriques : *toute suite possède au moins un point d'accumubtion.*

Nous renvoyons à l'article TOPOLOGIE GÉNÉRALE pour l'examen des très importantes propriétés des espaces compacts en mentionnant seulement ici le résultat suivant, de la théorie des espaces métriques, connu *sous* le nom de *théorème de Heine-Borel : une application continue d'un espace métrique compact dans un espace métrique est uniformément continue.* Donnons la démonstration qui est très simple et montrera comment on utilise la propriété de Borel-Lebesgue. Soit E et E' des espaces métriques de distances respectives d et d et f : E $\longrightarrow$E' une application continue. Choisissons un nombre réel positif ε. La continuité de f entraîne que, pour tout point $x \in$ E, il existe un nombre réel n(x), dépendant de x, tel que :

$$d(x,y) < \eta(x) \Rightarrow d'(f(x),f(y)) < \varepsilon ;$$

les boules ouvertes $B(x,\eta(x)/2)$, $x \in$ E, forment un recouvrement ouvert de E dont on peut, puisque E est compact par hypothèse, extraire un recouvrementjni fermé de boules de centres $x_1, x_2, \dots x_n$. Soit maintenant η le plus petit des nombres $\eta(x_1)/2, \dots \eta(x_n)/2$. Montrons que :

$$d(x,y) < \eta \Rightarrow d'(f(x),f(y)) < \varepsilon,$$

ce qui établira l'uniforme continuité de f. En effet, le point x appartient à au moins une des boules du sous-recouvrement fini, soit la boule de centre x_i, c'est-à-dire $d(x_i,x) < \eta(x_i)/2$, d'où :

$$d'(f(x_i),f(x)) < \frac{\varepsilon}{2},$$

puisque $d(x,y) < \eta(x_i)/2$, on a, par l'inégalité triangulaire, $d(x_i,y) < \eta(x_i)$, d'où :

$$d'(f(x_i),f(y)) < \frac{\varepsilon}{2},$$

et par suite, en utilisant encore une fois l'inégalité triangulaire,

$$d'(f(x),f(y)) \leqslant d'(f(x),f(x_i)) + d'(f(x_i),f(y)) < \varepsilon.$$

Les espaces numériques $\mathbf{R}^n$ ne sont pas compacts, mais tout point possède un système fondamental de voisinages compacts constitué par les boules fermées. On dit qu'un espace métrique dans lequel tout point a un système fondamental de voisinages compacts (ce qui revient à dire que, pour tout x, les boules fermées de centre x et de rayon assez petit sont compactes) est un *espace localement compact*.

3. Espaces métriques complets

Alors qu'au chapitre précédent les notions introduites (à l'exception de l'uniforme continuité ; cf. *infra*) sont topologiques, les notions de ce chapitre dépendent de manière essentielle de la distance.

Suites de Cauchy

B. Bolzano et A. Cauchy ont dégagé l'importance du critère de convergence suivant, qui ne fait pas intervenir la valeur de la limite : *une suite (u,) de nombres réels est convergente si et seulement si, pour tout 6 > 0, il existe un entier N tel que* :

$$p \geqslant N, q \geqslant N \Rightarrow |u_p - u_q| < \varepsilon.$$

(cf. CALCUL INFINITÉSIMAL Calcul à une variable). D'où le nom de *suite de Cauchy* donné à une suite (u_n) d'éléments d'un espace métrique E, de distance d, telle que, pour tout $\varepsilon > 0$, il existe un entier N tel que :

$$p \geqslant \mathbf{N}, q \geqslant \mathbf{N} \Rightarrow d(u_p,u_q) < \varepsilon.$$

Il est facile de voir que toute suite convergente (u_n), de limite a, est a fortiori une suite de Cauchy. Pour tout $\varepsilon > 0$, il existe N tel que $n \geqslant N \Rightarrow d(u_n, a) < \varepsilon/2$; pour $p \geqslant N$ et $q \geqslant N$, on a donc, en appliquant l'inégalité triangulaire, $d(u_p, u_q) \leqslant d(u_p, a) + d(a, u_q) < \varepsilon$. Au contraire, l'exemple de l'ensemble des nombres rationnels montre qu'une suite de Cauchy n'est pas toujours convergente. On dit qu'un espace métrique, comme $\mathbf{R}^n$, dans lequel toute suite de Cauchy est convergente, c'est-à-dire dans lequel les suites de Cauchy coïncident avec les suites convergentes, est un espace *complet*.

Les notions de suite de Cauchy et d'espace complet ne dépendent pas que de la topologie de l'espace métrique. Pour illustrer ce fait, désignons par R, l'ensemble des nombres réels muni de la distance usuelle, définie à partir de la valeur absolue, et par $\mathbf{R}_2$ ce même ensemble muni de la distance induite par celle de $\overline{\mathbf{R}}$ (cf. chap. 1) ; on a vu que les ouverts de $\mathbf{R}_1$ et $\mathbf{R}_2$ sont les mêmes (ce qui signifie que les deux distances sur R sont équivalentes). Considérons la suite des entiers naturels, dont le n-ième terme est l'entier n ; ce n'est manifestement pas une suite de Cauchy pour la distance usuelle sur R, mais c'est une suite de Cauchy dans $\mathbf{R}_2$, puisqu'elle converge vers l'élément $+\infty$ dans R. D'autre part, l'espace R, est complet, tandis que l'espace $\mathbf{R}_2$ ne l'est pas car la suite (de Cauchy) des entiers naturels n'est pas convergente dans cet espace : cela tient au fait que $\mathbf{R}_2$ n'est pas un sous-espace *fermé* de l'espace métrique $\overline{\mathbf{R}}$ (qui est complet car isométrique au segment $[-1, +1]$) ; plus généralement, la carac-

térisation des fermés au moyen des suites (cf. chap. 2) montre qu'un sous-espace métrique F d'un espace métrique complet est complet si et seulement s'il est fermé dans E.

Dans l'exemple précédent, l'application identique i : $\mathbf{R}_2 \to$ R,, définie par i(x) = x pour tout $x \in$ R, est continue, ce qui montre que l'image d'une suite de Cauchy par une application continue n'est pas nécessairement une suite de Cauchy. Par contre, les définitions montrent immédiatement que l'image *d'une suite de Cauchy par une application uniformément continue est* une *suite de Cauchy,* donc une suite convergente si l'espace d'arrivée est complet. Donnons, comme application de cette remarque, un important théorème de prolongement.

Théorème de prolongement des applications uniformément continues : Soit E un espace métrique, A un sous-ensemble partout dense de E (cela signifie, rappelons-le **que** tout point de E est adhérent à A) et F un espace métrique *complet.* Alors toute application f: A $\to$ F qui est *uniformément continue* se prolonge de manière unique en une application continue g : E $\to$ F ; de plus, g est uniformément continue.

Nous nous bornerons à indiquer l'idée de la démonstration. Pour $x \in$ A, on doit avoir g(x) $= f(x)$, et il faut donc définir g(x) pour $x \notin$ A. Or, puisque A est dense, il existe une suite (x_n) d'éléments de A qui converge vers x et, l'application g cherchée étant continue, on doit avoir nécessairement $g(x) = \lim g(x_n)$, avec $g(x_n) = f(x_n)$; mais la suite $(f(x_n))$, image d'une suite de Cauchy par une application uniformément continue, est une suite de Cauchy et, puisque F est complet, c'est une suite convergente, dont nous désignerons précisément la limite par g(x). Il reste à vérifier que g(s) ne dépend pas de la suite d'éléments de A qui converge vers x et que l'application g ainsi définie est uniformément continue, ce qui est un raisonnement élémentaire utilisant l'inégalité triangulaire.

Précisons enfin le lien entre les espaces métriques compacts et complets. Remarquons d'abord qu'une suite de Cauchy a au plus un seul point d'accumulation : ou bien elle n'en a aucun et elle ne converge pas, ou bien elle a exactement un point d'accumulation et elle converge alors vers ce point. Par suite, la propriété de Bolzano-Weierstrass montre que tout espace métrique compact est complet. Réciproquement, on montre qu'un espace métrique complet E est compact si et seulement s'il vérifie la propriété suivante de *précompacité* : Pour tout 6 > 0, il existe un recouvrement *fini* de E par des boules de rayon ε.

Complétion d'un espace métrique

La construction, due à Cantor, des nombres réels comme classes d'équivalence de suites de Cauchy de nombres rationnels (« suites fondamentales » dans la terminologie cantorienne) se transpose sans modification à un espace métrique quelconque.

Théorème de complétion. Pour tout espace métrique E, il existe un espace métrique complet $\hat{E}$ tel que E soit isométrique à un sous-espace partout dense de $\hat{E}$; de plus, l'espace $\hat{E}$ est déterminé à une isométrie près.

On dit que $\hat{E}$ **est** le *complété* de E, et on identifie dans la pratique E au sous-espace partout dense de $\hat{E}$ qui lui est isométrique. Ainsi, le complété de l'ensemble des rationnels pour la distance usuelle est l'ensemble des réels muni de la distance usuelle. Si on considère sur l'ensemble des rationnels la distance p-adique associée à un nombre premier *y,* on obtient comme complété l'ensemble des nombresp-adiques (cf. théorie des NOMBRES Nombres p-adiques).

Esquissons la construction du complété $\hat{E}$. Nous dirons que deux suites de Cauchy (x_n) et (y_n) d'éléments de E sont équivalentes si $d(x_n,y_n) \to 0$ pour $n \to \infty$. Il est clair que l'on définit ainsi une relation d'équivalence sur l'ensemble $\mathcal{E}$ de toutes les suites de Cauchy d'éléments de E ; nous désignerons par $\hat{E}$ le quotient de $\mathcal{E}$ par cette relation d'équivalence, et nous allons construire une distance sur $\hat{E}$. Si $x = (x_n)$ et $y = (y_n)$ sont deux suites de Cauchy, on montre que la suite de nombres réels $d(x_n, y_n)$ est convergente pour $n \to \infty$ (car c'est une suite de Cauchy) et que la limite ne change pas si on remplace x et y par des suites qui leur sont respectivement équivalentes ; cette limite ne dépend donc que des classes d'équivalence $\dot{x}$ et $\dot{y}$ des suites x et y et nous la désignerons par $\delta(\dot{x},\dot{y})$. On vérifie maintenant facilement que δ est une distance sur $\hat{E}$; le fait que $\hat{E}$ soit *complet* pour cette distance se montre par le célèbre « procédé diagonal » de Cantor, dont nous allons indiquer le principe : soit $(x^{(p)})$, $x^{(p)} = (x_1^{(p)}, \dots, x_n^{(p)}, \dots)$ une suite de Cauchy d'éléments de E (pour simplifier les notations, nous commettons ici l'abus de langage qui consiste à identifier $x^{(p)}$ avec sa classe d'équivalence) ; on montre alors que la suite $(x^{(p)})$ converge dans E vers la classe de la suite diagonale $(x_1^{(1)}, x_2^{(2)}, \dots, x_n^{(n)}, \dots)$.

Construisons maintenant une injection de E dans E. Pour $a \in E$, nous désignerons par $\varphi(a) \in E$ la classe d'équivalence de la suite constante (a, a, ... a, ...) ; il est clair que $\delta(\varphi(a),\varphi(b)) = d(a,b)$, et par suite φ réalise une isométrie de E sur le sous-espace métrique F de E formé des classes d'équivalence des suites constantes. L'ensemble F est partout dense dans E car si $x = (x_n)$ est une suite de Cauchy d'éléments de E, alors la suite $(\varphi(x_n))$ d'éléments de F converge vers $\dot{x}$ dans E.

La méthode des approximations successives

On doit à E. Picard une méthode de construction de solution d'équations par approximations successives (équations numériques, théorèmes d'existence et d'unicité d'équations différentielles ou intégrales ; cf. équations DIFFÉRENTIELLES, chap. 1 ; équations INTÉGRALES, chap. 2) que l'on peut formuler de la manière suivante dans le cadre des espaces métriques.

Théorème du point fixe. Soit E un espace métrique *complet* etfune application de E dans lui-même telle qu'il existe une constante k, $0 < k < 1$, avec $d(f(x),f(y)) \leqslant kd(x,y)$ quels que soient x et y dans E (on dit quefest une application contractante). Alors l'équation :

$$f'(x) = x$$

a une solution unique dans E. De plus, quel que soit $x_0 \in E$, la suite (x_n) définie par récurrence par $x_1 = f(x_0)$, $x_{n+1} = f(x_n)$ converge vers cette solution.

La démonstration est très simple et on voit clairement le rôle **joué** par les suites de Cauchy. Remarquons d'abord que l'unicité est évidente : si $f(x) = x$ et $f(y) = y$, on aura $d(f(x),f(y)) = d(x,y) \leqslant k\, d(x,y)$, d'où $d(x,y) = 0$. Il suffit donc de montrer que la suite (x_n) est convergente car, si sa limite est x, on obtient immédiatement $x = f(x)$ en faisant tendre n vers l'infini dans la relation de récurrence $x_{n+1} = f(x_n)$. Or on a :

$$d(x_{p+1},x_p) = d(f(x_p),f(x_{p-1})) \leqslant k\, d(x_p,x_{p-1}),$$

d'où, par récurrence sur p,

$$d(x_{p+1},x_p)\ 6\ k^p d(x_1,x_0)\,;$$

par l'inégalité triangulaire, on a maintenant, pour $q \geqslant p$,

$$\sum_{r=p}^{q-1} d(x_{r+1},x_r) \leqslant \left(\sum_{r=p}^{q-1} k^r\right) d(x_1,x_0).$$

soit, en remplaçant la progression géométrique de raison k par sa somme et en majorant cette dernière :

$$d(x_p,x_q) \leqslant \frac{k^p}{1-k} d(x_1,x_0),$$

et donc
$d(x_q,x_p) \to 0$ pour $p, q \to \infty$ puisque $0 < k < 1$. La suite de Cauchy (x_n) converge dans E vers une certaine limite x puisque E est complet. Remarquons que si on fait tendre y vers l'infini dans l'inégalité précédente, on obtient :

$$d(x,x_p) \leqslant \frac{k^p}{1-k} d(x_1,x_0),$$

qui précise la rapidité de convergence de la suite (x_n) vers x.

4. La propriété de Baire

Les sous-espaces ouverts des espaces métriques complets et les espaces métriques localement compacts possèdent la propriété suivante, appelée propriété *de Baire*, qui joue un rôle important dans de nombreuses questions d'analyse : Si U_n *est une suite d'ouverts partout denses, alors l'intersection des* U_n *est un ensemble partout dense*.

Sous le nom de « méthode de la catégorie », cc résultat a été utilisé systématiquement par l'école mathématique polonaise pour démontrer de profonds théorèmes (théorème du graphe fermé, théorème de Banach-Steinhaus ; cf. espaces vectoriels NORMÉS).

Par passage au complémentaire, on peut énoncer cette propriété : si F_n est une suite de fermés dont le complémentaire est partout dense, alors la réunion des F_n a un complémentaire partout dense. On dit qu'un ensemble A est *rare* si le complémentaire de son adhérence est partout dense ; ainsi, la propriété de Baire entraîne que, si un ensemble est réunion dénombrable d'ensembles rares (on dit qu'un tel ensemble est *maigre*), son complémentaire est partout dense. En particulier ce complémentaire est non vide ; ce type de raisonnement a été utilisé par S. Banach pour démontrer l'existence de fonctions possédant des singularités données à l'avance. Voici maintenant, pour terminer, un énoncé faisant intervenir la notion d'ensemble maigre et permettant, grâce à elle, de « dire quelque chose » d'une fonction qui est limite simple d'une suite de fonctions continues (on sait qu'une telle limite n'est pas nécessairement une fonction continue) : Si E est un espace métrique *complet* et F un espace métrique, soit f_n une suite d'applications continues de E dans F telles qu'en chaque point $x \in E$ la suite $(f_n(x))$ converge dans F vers un élément $f(x)$. Alors l'ensemble des points x de E où f n'est pas continue est un ensemble maigre.

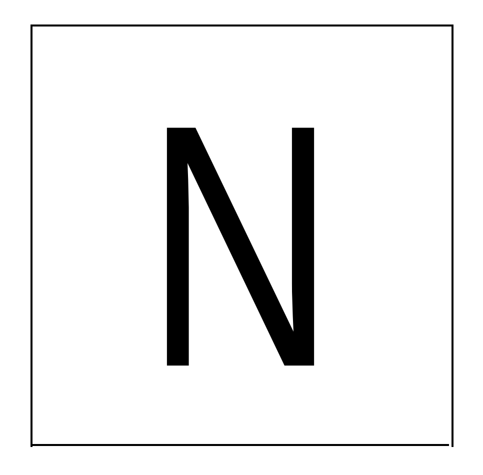
JEAN-LUC VERLEY

Bibliographie

N. BOURBAKI « Utilisation des nombres réels », in *Éléments de mathématiques*, liv. 3 : *Topologie générale*, Masson. nouv. éd., 1982 / G. CHOQUET, *Cours de topologie : espaces topologiques et espaces métriques, fonctions numériques, espaces vectoriels topologiques*, *ibid.*, 2' éd. 1984 / J. DIEUDONNÉ, *Fondements de l'analyse moderne*, Gauthier-Villars, Paris. 3' éd. 1979.

MULTILINÉAIRE ALGÈBRE

→ LINÉAIRE & MULTILINÉAIRE ALGÈBRE



NOMBRES THÉORIE DES

DANS la plupart des civilisations parvenues au stade de l'écriture, les nombres entiers ont, dès l'origine, été liés à des pratiques religieuses ou magiques, et leurs propriétés ont exercé une sorte de fascination sur les esprits, qui est loin d'être disparue de nos jours, où la « numérologie » conserve des adeptes ; il n'est donc pas étonnant que ce soit au sein de l'école pythagoricienne, imbue de mysticisme, qu'ait débuté l'étude scientifique de ces propriétés. Cette école entendait d'ailleurs mener de front les développements de la géométrie et de l'arithmétique en une « arithmogéométrie » où certains types de nombres étaient associés à des figures ; on associait par exemple, de façon assez naturelle, les nombres n^2 aux figures carrées : c'est ainsi que les pythagoriciens découvrent la formule $1 + 3 + 5 + ... + (2n - 1) = n^2$ en inscrivant dans un carré de côté n les carrés de côtés 1, 2, ... $n - 1$ ayant un même sommet. C'est aussi dans cette école, avec les « catégories » du pair et de l'impair, que commencent les réflexions sur la divisibilité : elles aboutissent, deux siècles plus tard, au magistral exposé d'Euclide. On sait qu'il démontre (aux notations près) l'existence et l'unicité de la décomposition d'un entier positif en facteurs premiers, et, par un raisonnement très ingénieux, l'existence d'une infinité de nombres premiers.

Aux pythagoriciens remontent également les premiers exemples d'équations diophantiennes, notamment la résolution de l'équation $p^2 + q^2 = r^2$ en nombres entiers ; c'était, dans leur « arithmogéométrie », la recherche des triangles rectangles à côtés commensurables. Diophante d'Alexandrie lui-même (sans doute au IVe siècle apr. J.-C.), s'il traite un grand nombre d'exemples de telles équations ou systèmes d'équations, ne s'intéresse en général qu'à la recherche de solutions en nombres *rationnels*, non nécessairement entiers. Mis à part quelques résultats isolés des mathématiques chinoise et indo-arabe sur des équations diophantiennes du premier et du second degré, la théorie des nombres ne recommence à se développer qu'avec Pierre de Fermat. Ses contributions portent à la fois sur la théorie de la divisibilité, avec le fameux théorème $a^{p-1} \equiv 1 \pmod{p}$ pour tout nombre premier p, et sur les équations diophantiennes, où on lui doit la première méthode générale d'attaque, la « descente infinie », dont le domaine d'application n'est pourtant pas défini avec précision et dépend avant tout de l'ingéniosité du mathématicien qui l'applique. Au XVIIIe siècle, L. Euler, J.-L. Lagrange et A.-M. Le Gendre, s'inspirant des idées de Fermat, prouvent la plupart des théorèmes seulement énoncés par ce dernier, et donnent en tout cas une solution complète pour les équations diophantiennes du second degré à deux inconnues ; c'est à ce propos qu'intervient une tech-

nique nouvelle, celle des fractions continuées, premier exemple d'utilisation d'approximations diophantiennes pour la résolution d'équations diophantiennes.

Jusque-la, les procédés de résolution d'équations diophantiennes consistaient en des manipulations algébriques élémentaires plus ou moins subtiles, pour permettre une application judicieuse de la théorie de la divisibilité des entiers rationnels. À partir du début du XIX[e] siècle, toutes les parties des mathématiques vont être progressivement mises à profit pour résoudre les problèmes de théorie des nombres.

Avec C. F. Gauss, développant des ébauches peu concluantes d'Euler, c'est d'abord l'extension de l'idée de divisibilité aux corps de nombres algébriques réels ou complexes ; il la développe en détail pour le corps Q(i) et, dans des notes non publiées de son vivant, pour certains corps cyclotomiques. Il faudra tout l'effort de l'école allemande du XIX[e] siècle (E. E. Kummer, L. Kronecker, R. Dedekind) pour surmonter, par la création de la théorie des idéaux, les difficultés provenant du fait que les anneaux d'entiers algébriques ne sont pas principaux en général (cf. la partie C ci-après - Nombres algébriques), et amener ainsi la théorie de la divisibilité dans ces anneaux au même point que la théorie d'Euclide pour Z. Le premier succès de cette nouvelle théorie est le critère obtenu par Kummer pour la non-résolubilité de l'équation de Fermat $x^n + y^n = z^n$, où n est premier, à savoir que n ne divise pas les numérateurs des $(n - 3)/2$ premiers nombres de Bernoulli.

C'est aussi avec **Gauss** que commence la théorie générale des formes quadratiques à coefficients entiers ; il traite en détail des formes binaires (dont la théorie équivaut à celle des corps quadratiques sur **Q**, comme devait le montrer plus tard Dedekind), et partiellement des formes ternaires. La théorie générale des formes quadratiques à n variables (théorèmes de réduction, de finitude et de représentations des formes les unes par les autres) est édifiée par F. G. Eisenstein, C. Hermite, H. J. S. Smith, H. Minkowski et H. Hasse, et prend sa forme quantitative générale avec les travaux de C. L. Siegel, qui utilise largement la théorie des fonctions analytiques et introduit à cette occasion la notion de fonction modulaire à n variables (cf. formes QUADRATIQUES). Les théorèmes de finitude sont d'autre part étendus par C. Hermite, C. Jordan et H. Poincaré aux formes de degré $\geqslant 3$.

Enfin, on sait que Gauss avait aussi découvert les fonctions elliptiques et la fonction modulaire (d'une variable), sans rien publier d'ailleurs sur ces sujets, et il n'avait pas manqué d'observer les liens entre cette théorie et certains problèmes de théorie des nombres. Pendant tout le XIX[e] siècle, l'exploration de ce nouveau domaine est menée avec vigueur, notamment par C. Jacobi, C. Hermite et L. Kronecker. Les résultats les plus profonds sont obtenus par ce dernier dans son étude de la multiplication complexe des fonctions elliptiques, qui le conduit au premier exemple de corps de classes.

P. G. Lejeune-Dirichlet, de son côté. inaugure deux nouvelles voies en théorie des nombres. D'abord, par sa « méthode des tiroirs », il montre comment traiter les questions d'approximations diophantiennes autrement que par la théorie des fractions continuées, et l'applique aussitôt pour obtenir un des théorèmes fondamentaux de la théorie des nombres algébriques, le théorème des unités (cf. la partie C ci-après - Nombres algébriques). Après lui, d'autres méthodes encore sont introduites dans la théorie des approximations diophantien-

nes : Hermite utilise à cette fin sa théorie des formes quadratiques et Minkowski sa « géométrie des nombres », obtenant des estimations remarquablement précises par des considérations très intuitives sur la position relative d'un ensemble convexe de $\mathbf{R}^n$ par rapport au réseau $\mathbf{Z}^n$ (cf. approximations DIOPHANTIENNES). J. Liouville, de son côté, obtenant le premier théorème de « non-approximation » des nombres algébriques, donne le premier exemple effectif de nombres transcendants, et Hermite, en combinant habilement des méthodes d'approximation diophantienne à la théorie des fonctions analytiques d'une variable complexe, prouve en 1872 la transcendance du nombre e, résultat spectaculaire que F. Lindemann complète, dix ans plus tard, en établissant, par une méthode analogue, la transcendance de π.

L'autre contribution fondamentale de Dirichlet est l'introduction des séries qui portent son nom et dont il se sert pour démontrer le théorème de la progression arithmétique, ainsi que pour obtenir une expression explicite du nombre de classes d'idéaux d'un corps quadratique (cf. la partie A ci-après - Théorie analytique des nombres). Ces deux types d'application des séries de Dirichlet vont être considérablement développés : ils conduisent, d'une part, avec B. Riemann, J. Hadamard et C. de La Vallée-Poussin à la démonstration du théorème des nombres premiers et, d'autre part, avec Dedekind, aux généralisations des fonctions L à tous les corps de nombres algébriques. préludes à des généralisations ultérieures encore plus vastes (cf. fonction ZÊTA).

Le développement de la théorie des nombres, depuis la dernière décennie du XIX[e] siècle, se caractérise par l'apport de plus en plus important de deux théories qui jusqu'alors n'y avaient joué qu'un rôle secondaire et implicite : la géométrie algébrique et la théorie des groupes. Grâce aux puissants moyens empruntés à ces théories, on dispose pour la première fois de quelques théorèmes *généraux* sur les équations diophantiennes ; bien que l'on soit encore très loin d'une compréhension profonde des phénomènes étudiés, on peut espérer être sur la bonne voie et arriver un jour à une théorie unique qui engloberait à la fois ce que nous appelons maintenant la théorie des groupes algébriques, celle des « groupes arithmétiques » et une bonne part de la théorie des nombres.

Avant de donner quelques indications sur ces développements, mentionnons que l'approfondissement des méthodes héritées du XIX[e] siècle a aussi permis d'importants progrès : l'extension de la méthode d'Hermite pour la preuve de la transcendance de e a conduit, par une série d'améliorations successives (Siegel, Gelfond), aux résultats généraux de A. Baker sur les nombres transcendants (cf. nombres TRANSCENDANTS). Une évolution analogue pour les approximations diophantiennes (A. Thue, Siegel, Dyson) a de même permis d'améliorer le résultat initial de « non-approximation » de Liouville : le meilleur résultat (K. Roth) est que, pour un nombre algébrique non rationnel α et p, q entiers ($q \geqslant 1$), on a, pour tout $\varepsilon > 0$,

$$(1) \qquad \left|\alpha - \frac{p}{q}\right| \geqslant \frac{c(\varepsilon)}{q^{2+\varepsilon}},$$

mais la démonstration de ce résultat ne donne aucun moyen de calculer effectivement $c(\varepsilon)$ pour α et ε donnés. Baker, par ses méthodes, a obtenu la beaucoup moins bonne inégalité :

$$(2) \qquad \left|\alpha - \frac{p}{q}\right| \geqslant cq^{-d} \exp ((\ln q)^{1/k}),$$

où $d \geqslant 3$ est le degré de α et $k > d+1$, c étant cette fois une constante $c(\alpha, k)$ que l'on peut minorer explicitement en fonction de α et k.

A. Thue avait déjà déduit de son théorème de « non-approximation », par un raisonnement très ingénieux, le fait que, si $P(x, y)$ est un polynôme à deux variables, à coefficients entiers, *homogène* et de degré $\geqslant 3$, alors l'équation $P(x, y) = m$, où m est entier $\neq 0$, ne peut avoir qu'un nombre *fini* de solutions (x, y) en nombres entiers. Ce théorème a été généralisé et mis sous sa forme définitive par Siegel en 1929 : Pour tout polynôme (homogène ou non) $P(x, y)$ à coefficients entiers, tel que le *genre* de la courbe C d'équation $P(x, y) = 0$ soit $\geqslant 1$, il ne peut y avoir sur C qu'un nombre *fini* de points (x, y) à coordonnées entières. Ce résultat fait déjà intervenir une notion profonde de géométrie algébrique (le genre d'une courbe) dans sa formulation. Sa démonstration combine le théorème de « non-approximation » d'un nombre algébrique avec le théorème de Mordell-Weil. Ici encore, la formulation même de ce théorème fait intervenir la géométrie algébrique, et sa démonstration utilise un raisonnement de « descente infinie » sur la jacobienne de C (variété abélienne de dimension égale au genre de C). L. J. Mordell a conjecturé que, sous les hypothèses du théorème de Siegel, il n'y a même qu'un nombre fini de points de C à coordonnées rationnelles lorsque le genre de C est $\geqslant 2$. Cette conjecture a été démontrée en 1983 par G. Faltings ; cette démonstration ouvre un nouveau chapitre de la théorie des nombres (cf. équations DIOPHANTIENNES). D'autre part, Baker, par ses méthodes, a pu dans certains cas améliorer le théoréme de Siegel en donnant une majoration effective pour les solutions entières de $P(x, y) = m$; par exemple, pour l'équation $y^2 = x^3 + D$, avec $D \neq 0$, on a nécessairement : $\sup(|x|, |y|) < \exp(10^{10}|D|^{10^4})$.

Une autre source de progrès en théorie des nombres est la géométrie algébrique sur un *corps fini*, développée par E. Artin, par H. Hasse et surtout par A. Weil. Ainsi une congruence algébrique telle que $P(x_1, ..., x_r) \equiv 0 \pmod p$, où p est premier, est une *équation* algébrique $P(\bar{x}_1, \dots, \bar{x}_r) = 0$ entre les classes $\bar{x}_j$ modulo p des entiers x_j, autrement dit entre éléments du *corps fini* $\mathbf{Z}/p\mathbf{Z} = \mathbf{F}_p$. Cette interprétation est une des raisons qui ont conduit à développer la géométrie algébrique sur un corps de base autre que les corps usuels R ou C (cf. GÉOMÉTRIE ALGÉBRIQUE). Il y a en particulier une parenté très étroite entre les corps de nombres algébriques et les corps de fonctions rationnelles sur une courbe d'équation $P(x, y) = 0$ à coefficients dans un corps fini ; toute la théorie de la divisibilité se développe de façon semblable dans les deux cas (cf. *infra*), et on peut aussi définir les « fonctions L » et la fonction zêta de la même manière pour ces deux types de corps. Le résultat fondamental obtenu par Weil est que, pour ces dernières « fonctions zêta » (qui sont ici rationnelles en p^{-s}), l'« hypothèse de Riemann » est vraie ; cela entraîne des majorations pour les sommes d'exponentielles qui jouent un grand rôle en théorie analytique des nombres (cf. la partie A ci-après - Théorie analytique des nombres) : par exemple, si $f(x)$ est un polynôme sur $\mathbf{F}_p$ sans facteur carré et de degré impair m, on a l'inégalité (qu'on ne sait pas prouver par d'autres méthodes) :

$$(3) \qquad \left| \sum_{x \in \mathbf{F}_p} \left(\frac{f(x)}{p} \right) \right| \leqslant (m-1)\sqrt{p},$$

où $\left(\frac{n}{p}\right)$ est le symbole de Le Gendre. Au moyen de cette inégalité, on prouve par

exemple que le plus petit non-résidu quadratique modulo p tend vers l'infini moins vite que $p^{1/4\sqrt{e}+\varepsilon}$, $p \to \infty$, quel que soit $\varepsilon > 0$.

Le passage d'une équation diophantienne $P(x_1, \ldots, x_r) = 0$ à la congruence modulo p correspondante $P(x_1, \ldots, x_r) \equiv 0 \pmod{p}$ est un des procédés les plus anciens d'étude de ces équations. Mais on peut aussi considérer les congruences $P(x_1, \ldots, x_r) \equiv 0 \pmod{p^n}$ pour tous les exposants n, et, en un certain sens, l'existence de solutions *pour tout* n est un résultat « proche » de l'existence de solutions de l'équation initiale (la différence de deux entiers peut être non nulle et divisible par une puissance élevée de p, mais non par *toutes* les puissances de p). En approfondissant cette idée, on a été conduit à l'introduction des nombres p-adiques de Hensel et à leurs généralisations : ils fournissent un procédé systématique pour « localiser » en un nombre premier (resp. un idéal premier) l'étude de l'anneau des entiers Z (resp. d'un anneau d'entiers algébriques), procédé tout semblable au développement en série entière au voisinage d'un point d'une fonction rationnelle sur une courbe algébrique (cf. la partie B ci-après - Nombres p-adiques) ; on touche là à une nouvelle et profonde influence de la géométrie algébrique, qui n'a cessé d'inspirer Hensel et ses successeurs.

L'étude de la divisibilité dans les corps de nombres algébriques se conçoit maintenant comme une synthèse « globale » des propriétés « locales » d'un tel corps en toutes ses « places » ; les outils essentiels pour en exprimer les résultats sont les groupes d'*adèles* et d'*idèles* introduits par C. Chevalley (cf. la partie C ci-après - Nombres algébriques). Ces groupes sont naturellement munis de topologies, qui en font des groupes commutatifs localement compacts ; et on constate alors que les théorèmes fondamentaux sur la divisibilité dans les corps de nombres algébriques s'interprètent de façon extrêmement simple et frappante en langage de la théorie des groupes topologiques.

Mais ce n'est pas tout : en premier lieu, on a de même une interprétation remarquable dans ce langage des résultats fondamentaux de la théorie du corps de classes. Déjà, sous l'impulsion de D. Hilbert, cette théorie avait été rattachée à la théorie de Galois (cf. CORPS) ; le problème central en est la détermination du groupe de Galois de l'extension abélienne maximale A, d'un corps de nombres algébriques k. Or, la théorie des idèles en donne une expression explicite : si I_k est le groupe des idèles de k, P_k le sous-groupe (discret) des idèles principaux, $C_k = I_k/P_k$ le groupe (compact) des classes d'idèles et D_k la composante neutre de C_k, le groupe de Galois de A, sur k est canoniquement isomorphe à C_k/D_k. On a des résultats tout à fait analogues pour la divisibilité et la théorie du corps de classes lorsque k est un corps de fonctions rationnelles sur une courbe algébrique définie sur un corps fini.

En second lieu, toute la théorie de l'analyse harmonique générale (cf. analyse HARMONIQUE) est applicable aux groupes d'adèles et d'idèles ; le développement de cette théorie a conduit Tate, Ono, Tamagawa et A. Weil à insérer dans ce nouveau cadre les travaux de Hecke sur les fonctions L et ceux de Siegel sur les formes quadratiques (cf. fonction ZÊTA et formes QUADRATIQUES). Pour aller plus loin et « sortir du commutatif », la voie qui semble aujourd'hui la plus prometteuse est la théorie des groupes de Lie (généralisée aux groupes algébriques, p-adiques et « adéliques » de façon convenable), et notamment celle des représentations de ces groupes dans des espaces fonctionnels

convenables (cf. GROUPES - Groupes de Lie). Déjà, on a subordonné à cette théorie les résultats classiques d'Hermite, de Jordan et de Minkowski sur la « réduction » des formes (A. Bore1 et Harish-Chandra) et la théorie arithmétique des fonctions automorphes, notamment celle des formes modulaires de Hecke, Siegel, Maass et Petersson (Shimura et Jacquet-Langlands). Il y a lieu d'espérer qu'on aboutira par cette voie à une généralisation satisfaisante de la théorie du corps de classes pour les extensions galoisiennes non abéliennes des corps de nombres algébriques ; peut-être aussi ces méthodes permettront-elles de comprendre les raisons du succès de méthodes de théorie analytique des nombres telles que la méthode de Hardy-Littlewood (cf. la partie A ci-après - Théorie analytique des nombres).

JEAN DIEUDONNÉ



A. Théorie analytique des nombres

Ce qu'on appelle la « théorie analytique des nombres » ne peut pas être considéré comme une théorie mathématique au sens usuel qu'on donne à ces mots, c'est-à-dire un système organisé de définitions et de théorèmes généraux accompagné d'applications à des exemples importants. Il s'agit au contraire ici presque exclusivement de problèmes particuliers qui se posent en arithmétique et qui, pour la plupart, consistent à étudier (cf. calculs ASYMPTOTIQUES pour la position du problème et les notations o et 0 de Landau) l'« allure à l'infini » de certainesfonctions définies par des conditions de nature arithmétique : par exemple le nombre $\pi(x)$ de nombres premiersp $\leqslant x$ ou le nombre U(n) des solutions de l'équation $x_1^2 + x_2^2 = n$ en nombres entiers. Depuis 1830, on a imaginé, pour résoudre ces questions, des méthodes d'une extraordinaire ingéniosité qui consistent à associer aux fonctions arithmétiques étudiées des *fonctions analytiques* auxquelles on peut appliquer la théorie de Cauchy ou l'analyse harmonique ; mais, malgré les succès spectaculaires obtenus par ces méthodes, on ne peut dire que l'on en comprenne vraiment les raisons profondes.

1. La théorie additive

Le point de vue formel

Un *monoïde* est un ensemble M où est définie une loi de composition $(s, t) \mapsto st$ qui est *associative* et possède un élément neutre e (autrement dit $es = se = s$ pour tout $s \in$ M) ; les groupes sont évidemment des monoïdes ; d'autres exemples importants sont formés par l'ensemble N des entiers $\geqslant 0$, avec pour loi l'addition, et l'ensemble N* des entiers > 0, avec pour loi la multiplication. Étant donné un corps commutatif K, on définit, pour tout monoïde M, l'*algèbre* K[M] du *monoïde* M *sur* K de la façon suivante : on définit l'espace vectoriel K[M] à l'aide d'une hase (u_s), dite canonique, où l'ensemble d'indices est M ; puis on prend pour table de multiplication de cette base $u_s u_t = u_{st}$, quels que soient s et t dans M ; on vérifie qu'on a bien défini ainsi une algèbre associative dont l'élément unité est u_e. Tout élément $x \in$ K[M] s'écrit d'une seule manière :

$$\sum \xi_s u_s,$$

avec $\xi_s \in$ K et $\xi_s = 0$ sauf pour un nombre *fini* de valeurs de $s \in$ M ; il revient au même de dire que K[M] est formé des familles (ξ_s), $s \in$ M, d'éléments de K,

indexées par M, telles que $\xi_s = 0$ sauf pour un nombre fini d'éléments de M ; l'addition se fait composante par composante et la multiplication est définie par :

$$(1) \qquad (\xi_s)(\eta_s) = (\zeta_s),$$

avec :

$$(2) \qquad \zeta_s = \sum_{vw = s} \xi_v \eta_w,$$

somme qui a un sens dans K, puisqu'elle n'a qu'un nombre *jîni* de termes $\neq 0$.

Remarquons maintenant que, si l'on ne fait aucune hypothèse sur les familles (ξ_s) et (η_s), le second membre de (2) a encore un sens si le monoïde M satisfait à la condition : (D) *Pour tout* $s \in$ M, *il n'existe qu'un nombre fini de couples* (v, w) *d'éléments de* M *tels que* $vw = s$.

Par exemple, les monoïdes N et N* définis ci-dessus vérifient (D). Pour un tel monoïde, on définit donc sur l'espace vectoriel K[[M]] de *toutes* les familles (ξ_s), $s \in$ M, une structure d'algèbre par les formules (1) et (2) ; on dit que cette algèbre est l'*algèbre large* du monoïde M.

Lorsque M = N, K[[N]] n'est autre que *l'algèbre des séries formelles à* une indéterminée : si l'on pose u_1 = X, on a $u_n = X^n$ pour tout entier n $\geqslant 1$; au lieu d'écrire (ξ_n), $n \in$ N, les éléments de cette algèbre, on convient de les noter :

$$\sum_{n=0}^{\infty} \xi_n X^n,$$

la loi de multiplication (2) donnant alors la formule usuelle :

$$(2') \quad \Big(\sum_{n=0}^{\infty} \xi_n X^n\Big)\Big(\sum_{n=0}^{\infty} \eta_n X^n\Big) = \sum_{n=0}^{\infty} \Big(\sum_{p+q=n} \xi_p \eta_q\Big) X^n$$

du produit de séries entières. On note encore cette algèbre K[[X]]. Elle contient évidemment l'algèbre des polynômes K[X] ; en outre, pour qu'une série formelle :

$$\sum_{n=0}^{\infty} \xi_n X^n$$

ait un *inverse* dans K[[X]], il faut et il suffit que $\xi_0 \neq 0$. Les fractions rationnelles P(X)/Q(X) telles que Q(0) $\neq 0$ sont donc des éléments de K[[X]] ; en particulier, on a :

$$(3) \quad \frac{1}{1 - x} = 1 + x + X^2 + \quad + X^n +$$

L'ordre $\omega(f)$ d'une série formelle :

$$f(X) = \sum_{n=0}^{\infty} \xi_n X^n$$

non nulle est le plus petit exposant n tel que $\xi_n \neq 0$. Soit (f_n) une suite de séries formelles telle que l'ordre $\omega(f_n)$ tende vers $+\infty$ avec n. Alors, on peut définir dans K[[X]] la *somme infinie :*

$$(4) \qquad f_1 + f_2 + \quad + f_n +$$

et le *produit infini :*

$$(5) \qquad (1 + f_1)(1 + f_2)...(1 + f_n)...$$

de la façon suivante. Pour tout entier m, il existe un entier N(m) tendant vers $+\infty$ avec m tel que, dans la somme $f_1 + f_2 + \quad + f_n$, resp. le produit $(1 + f_1)...(1 + f_n)$, tous les termes de degrés $\leqslant m$ soient les mêmes dès que $n \geqslant N(m)$. Il y a donc une série formelle h(X) et une seule telle que, pour tout entier m, les termes de degrés $\leqslant m$ dans h(X) soient les mêmes que ceux des sommes $f_1 + f_2 + \quad + f_n$, resp. des produits $(1 + f_1)...(1 + f_n)$, pour tout $n \geqslant N(m)$; c'est cette série qui, par définition, est la

somme infinie (4), resp. le produit infini (5). Cette définition justifie, pour une série formelle, l'écriture :

$$\sum_{n=0}^{\infty} \xi_n X^n.$$

On supposera toujours par la suite que K est le corps C des nombres complexes ; si (ξ_n) est une suite de nombres complexes, on dit souvent que la série formelle :

$$\sum_{n=0}^{\infty} \xi_n X^n$$

est la *série génératrice* de la suite (ξ_n). Cela étant, des propriétés d'une suite (ξ_n) on peut souvent déduire une expression de la série génératrice sous une autre forme qui permet d'obtenir des relations entre (ξ_n) et d'autres suites.

Exemples

Désignons par ξ_n le nombre de solutions en *entiers* $\geqslant 0$ de l'équation diophantienne à trois variables $x + 2y + 3z = n$. En raison de (3), la série génératrice :

$$\sum_{n=0} \xi_n X^n$$

est égale à :

$$\frac{1}{(1-X)(1-X^2)(1-X^3)}.$$

Si l'on décompose cette fraction rationnelle en éléments simples, on obtient :

$$\frac{1}{6(1-X)^3} + \frac{1}{4(1-X)^2} + \frac{17}{72(1-X)} + \frac{1}{8(1+X)} + \frac{1}{9(1-jX)} + \frac{1}{9(1-j^2X)},$$

où j désigne la racine cubique de l'unité $\exp(2\pi i/3)$; utilisant de nouveau (3) ainsi que les formules analogues pour $1/(1 - X)^k$, qui s'obtiennent en dérivant un nombre quelconque de fois les deux membres de (3), on obtient aisément l'expression de ξ_n comme *l'entier le plus proche de* $(n+3)^2/12$.

Plus généralement, soit $a_1, a_2, \ldots, a_r$ des entiers > 0, sans diviseur commun $\neq 1$, et notons ξ_n le nombre de solutions en *entiers* $\geqslant 0$ de l'équation à r inconnues :

(6) $$a_1x_1 + a_2x_2 + \ldots + a_rx_r = n.$$

On voit que la série génératrice correspondante est :

$$\frac{1}{(1-X^{a_1})(1-X^{a_2})\ldots(1-X^{a_r})}.$$

La décomposition de cette fraction rationnelle en éléments simples donne encore ξ_n ; si l'on ne cherche qu'une partie principale de ξ_n, on constate aisément qu'elle provient du pôle d'ordre le plus élevé, c'est-à-dire le point 1, et l'on trouve :

$$\xi_n \sim n^{r-1}/(a_1a_2\ldots a_r(r-1)!).$$

Sommes de carrés

Il peut se faire que la série génératrice d'une suite (ξ_n) de nombres complexes fournisse une série entière *convergente* dans un voisinage de $z = 0$ lorsqu'on y substitue à l'indéterminée X un nombre complexe z. Des propriétés de la fonction analytique $f(z)$ égale à la somme de cette série entière, on peut alors déduire des renseignements sur la valeur de ξ_n.

Les premiers exemples de cette méthode ont été donnés par Jacobi à l'aide de sa théorie des fonctions elliptiques, pour le problème consistant à chercher le nombre de solutions en nombres *entiers* (positifs ou négatifs) de l'équation à r inconnues :

(7) $$x_1^2 + x_2^2 + \ldots + x_r^2 = n.$$

Ce nombre est le coefficient de z^n dans le développement en série entière de la fonction $(f(z))^r$, où :

$$(8) \qquad f(z) = \sum_{m=-\infty}^{\infty} z^{m^2},$$

série entière qui converge pour $|z| < 1$. Or f(z) est une des « fonctions thêta » de la théorie des fonctions elliptiques, et certaines de ses puissances s'expriment par d'autres développements en série qui fournissent le nombre de solutions de (7).

Par exemple, pour $r = 2$, on a :

$$(f(z))^2 = 1 + 4 \sum_{n=1}^{\infty} \frac{z^n}{1+z^{2n}}$$

$$= 1 + 4 \sum_{n=1}^{\infty} \frac{z^n - z^{3n}}{1 - z^{4n}},$$

d'où l'on déduit que le nombre U(n) de solutions de $x_1^2 + x_2^2 = n$ est *quatre fois la différence entre le nombre des diviseurs de n de lu forme $4k+1$ et le nombre des diviseur.~ de n de la forme $4k+3$.*

De même, pour $r = 4$, on a la formule :

$$f(z)^4 = 1 + 8\left(\frac{z}{1-z} + \frac{2z^2}{1+z^2} + \frac{3z^3}{1-z^3} + \dots\right)$$

$$= 1 + 8 \sum_m \sigma_1(m)(z^m + 3z^{2m} + 3z^{4m} + 3z^{8m} + \dots).$$

où m parcourt l'ensemble des nombres impairs $\geqslant 1$ et où $\sigma_1(m)$ désigne la somme des diviseurs de m. On en conclut que, si l'on pose $n = 2^\alpha m$, où m est impair, le nombre de solutions de :

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = n$$

est $8\,\sigma_1(m)$ si $\alpha = 0$ et $24\,\sigma_1(m)$ si $\alpha > 0$; en particulier ce nombre est toujours $\geqslant 1$ (théorème de Lagrange).

Il convient de noter ici qu'il y a des formules explicites pour le nombre de solutions de (7) pour $r \leqslant 8$ qui se déduisent de la théorie générale des formes quadratiques à coefficients entiers, cette théorie « expliquant » aussi l'absence de telles formules pour $r > 8$ (cf. formes QUADRATIQUES).

Partitions

Le « nombre de partitions » $p(n)$ d'un entier $n \geqslant 1$ est par définition le nombre de solutions en *entiers* $\geqslant 0$ de l'équation :

$$(9) \qquad x_1 + 2x_2 + 3x_3 + \dots + mx_m + \dots = n,$$

où le nombre d'inconnues x_m n'est pas limité (mais, pour un n donné, il est clair qu'on a nécessairement $x_m = 0$ pour tout $m > n$). Ce nombre peut aussi être défini comme le nombre des classes d'équivalence des *partition.~* d'un ensemble de n éléments, lorsqu'on range dans une même classe deux partitions qui se déduisent l'une de l'autre par une permutation de l'ensemble. Il est immédiat que la série génératrice converge pour $|z| < 1$ et est donnée par :

$$(10) \qquad F(z) = \sum_{n=0}^{\infty} p(n) z^n = \prod_{m=1}^{\infty} (1 - z^m)^{-1}.$$

L'idée fondamentale est d'exprimer le coefficient p(n) à l'aide de la formule de Cauchy :

$$(11) \qquad p(n) = \frac{1}{2\pi i} \int_C \frac{F(x)}{x^{n+1}} dx,$$

où C est un cercle de centre 0 et de rayon $r < 1$; le problème est d'évaluer cette intégrale lorsque r est pris voisin de 1. On utilise le fait que $F(z)$ est liée à une « fonction modulaire » $\eta(t)$ par la formule :

$$(12) \qquad F(e^{2\pi i t}) = e^{\pi i t/12}(\eta(t))^{-1},$$

où t est un nombre complexe de partie imaginaire strictement positive. La propriété essentielle de cette fonction η est la formule de transformation :

(13) $$\eta(s.t) = \varepsilon_s(ct + d)^{1/2}\eta(t),$$

pour toute transformation modulaire :

(14) $$s : t \mapsto \frac{at+b}{ct+d},$$

où a, b, c, d sont des entiers tels que $ad - bc = 1$ et où ε_s est une racine 24-ième de l'unité, déterminée explicitement en fonction de a, b, c, d. Comme il est immédiat que $z = 1$ est un point singulier de la fonction F, il en est de même, en vertu de (12) et (13), de tout point « rationnel » sur le cercle unité, c'est-à-dire de la forme $\exp(2\pi ih/k)$ où h et k sont des entiers premiers entre eux. La méthode extrêmement ingénieuse de Hardy-Ramanujan, perfectionnée par Rademacher, consiste à faire intervenir les contributions de ces points singuliers dans l'intégrale (11), en divisant le cercle C en arcs partiels dont chacun est « voisin » d'un tel point. De façon précise, pour tout entier N, on considère l'intégrale (11) étendue au cercle C de rayon $\exp(-2\pi/N^2)$. On range en une suite croissante les fractions h/k à termes premiers entre eux et tels que $0 \leqslant h \leqslant k \leqslant N$ (*suite de Farey* d'ordre N). La théorie élémentaire de la suite de Farey conduit, pour trois fractions consécutives $h'/k' < h/k < h''/k''$, à « isoler » la fraction centrale par l'intervalle $I_{h,k}$ d'extrémités :

$$\frac{h+h'}{k+k'}, \quad \frac{h+h''}{k+k''};$$

on démontre que ces extrémités s'écrivent aussi :

$$\frac{h}{k} - \frac{1}{k(k+k')}, \quad \frac{h}{k} + \frac{1}{k(k+k'')},$$

les longueurs :

$$L_1 = \frac{1}{k(k+k')}, \quad L_2 = \frac{1}{k(k+k'')}$$

étant donc bornées par :

$$\frac{1}{k(2N-1)} \leqslant L_i \leqslant \frac{1}{k(N+1)},$$

pour $i = 1, 2$, ce qui permet les majorations ultérieures. À ces intervalles correspondent sur le cercle C les arcs $\xi_{h,k}$ de la « dissection de Farey » définis par :

$$x = \exp\left(-\frac{2\pi}{N^2} + 2\pi i\theta\right), \quad \theta \in I_{h,k}.$$

Utilisant (14), on est conduit à remplacer dans l'intégrale :

$$\int_{\xi_{h,k}} F(x)x^{-n-1}\,dx$$

la fonction F(x) par l'expression :

$$\omega_{h,k}\Psi_k(z),$$

avec :

$$x = \exp\left(2\pi i\left(\frac{h}{k} + \frac{iz}{k}\right)\right),$$

de sorte que :

$$z = k\left(\frac{1}{N^2} - i\varphi\right),$$

où φ varie de $-L_2$ à L_1, où Ψ_k est la fonction « élémentaire » :

$$\Psi_k(z) = z^{1/2}\exp\left(\frac{\pi}{12k}\left(\frac{1}{z} - z\right)\right),$$

et enfin où $\omega_{h,k}$ est la racine 24-ième de l'unité définie par :

$$\omega_{h,k} = \exp(\pi i S(h,k)),$$

avec :

$$S(h,k) = \sum_{j=0}^{k-1}\left(\frac{hj}{k} - \left[\frac{hj}{k}\right] - \frac{1}{2}\right)\left(\frac{j}{k} - \frac{1}{2}\right)$$

Un calcul assez élémentaire de majoration montre que, lorsqu'on fait ces substitutions dans les diverses intégrales étendues aux arcs $\xi_{h,k}$, l'erreur totale commise est $O(N^{-1/2})$. Il reste à évaluer les intégrales :

$$\int_{\xi_{h,k}} \Psi_k(z) x^{-n-1}\, dx\,;$$

cela se fait à l'aide d'une utilisation très ingénieuse de la théorie de Cauchy et conduit finalement au remarquable résultat suivant obtenu par H. Rademacher (1937). Le nombre p(n) s'exprime par une *série convergente* :

$$(15)\quad p(n) = \frac{1}{\pi\sqrt{2}} \sum_{k=1}^{\infty} A_k(n)\sqrt{k}\,\frac{d}{dn}\left(\frac{\operatorname{sh}(cu(n)/k)}{u(n)}\right),$$

où sh désigne le sinus hyperbolique, ou $c = \pi\sqrt{2/3}$, où :

$$A_k(n) = \sum_{(h,k)=1} \omega_{h,k} \exp(-2\,\pi i n h/k),$$

h parcourant l'ensemble des entiers $< k$ premiers à k, **et** enfin où :

$$u(n) = \left(n - \frac{1}{24}\right)^{1/2};$$

on peut obtenir des majorations commodes du *reste* de cette série en fonction de M lorsqu'on ne considère que les M premiers termes. Pour obtenir la valeur de $p(n)$, il suffit évidemment de prendre M tel que ce reste soit en valeur absolue $< 1/2$ puisque p(n) est un entier. Par exemple, pour $n = 721$, il suffit de prendre M = 21, et l'on trouve :

$$p(721) = 161061755750279477635534762.$$

De plus, chacun des termes de la série (15), considéré comme fonction de n, est négligeable devant le précédent ; en particulier, on a la partie principale (résultat dû à G. H. Hardy et à S. Ramanujan) :

$$(16)\quad p(n) \sim \frac{1}{4\sqrt{3}n} \exp\left(\pi\sqrt{\frac{2}{3}}\, n^{1/2}\right),\quad n \to \infty.$$

Signalons encore que la formule (10) a permis à Ramanujan, en se plaçant au point de vue formel, d'obtenir des propriétés arithmétiques intéressantes des nombres p(n) ; par exemple, on a les congruences :

$$p(5n+4) \equiv 0 \pmod{5},$$
$$p(7n+5) \equiv 0 \pmod{7},$$
$$p(25n+24) \equiv 0 \pmod{25}.$$

Le problème de Waring

Le théorème de Lagrange sur la possibilité d'écrire tout entier comme somme de quatre carrés au plus a conduit en 1792 le mathématicien anglais E. Waring à avancer la conjecture que, pour tout exposant entier $k \geqslant 2$, il existe un entier $g(k)$ tel que l'équation :

$$(17)\qquad x_1^k + x_2^k + \dots + x_{g(k)}^k = n$$

possède au moins une solution en nombres entiers, quel que soit l'entier $n \geqslant 0$. La première démonstration de cette conjecture fut donnée par Hilbert en 1909 ; on dispose actuellement de méthodes beaucoup plus puissantes, dues à G. H. Hardy, à J. E. Littlewood et à I. M. Vinogradov et qui non seulement prouvent la conjecture de Waring avec une bonne estimation de $g(k)$, mais encore donnent une estimation approchée du *nombre de solutions* de (17) en nombres entiers. L'idée de Hardy et de Littlewood est de généraliser la méthode de Jacobi en considérant la fonction :

$$f(z) = \sum_{m=0}^{\infty} z^{m^k},$$

série entière convergente pour $|z| < 1$. Le nombre de solutions de :

$$(18) \qquad x_1^k + \dots + x_s^k = n$$

en entiers $\geqslant 0$ est donc donné, comme dans (11), par la formule de Cauchy :

$$(19) \qquad r_0(n) = \frac{1}{2\pi i}\int_C (f(x))^s x^{-n-1}\,dx,$$

l'intégrale étant étendue à un cercle de centre 0 et de rayon $r < 1$. La fonction f n'a ici aucune propriété analogue à (12) et (13), mais $z = 1$ en est évidemment un point singulier, et des considérations heuristiques montrent qu'il en est encore de même de tout point « rationnel » $\exp(2\pi ip/q)$, la « contribution » d'un tel point dans l'intégrale (18) étant d'autant plus importante que q est plus petit. La *méthode de Hardy-Littlewood* consiste à décomposer encore le cercle C en arcs partiels centrés aux points correspondant à une suite de Farey ; on ne peut plus ici obtenir de majoration commode pour les arcs $\xi_{h,k}$ tout entiers ; il faut les restreindre d'une certaine manière, obtenant ce qu'on appelle les « *major arcs* » ; leur contribution donne la partie principale de $r_0(n)$; mais. pour le prouver, il faut majorer la contribution des arcs restants de C (« *minor arcs* »). Les calculs de majoration et d'approximation sont ici beaucoup plus difficiles que pour le problème des partitions.

On peut, avec Vinogradov, présenter cette méthode d'une manière un peu différente ; pour un n donné, on a évidemment $|x_j| \leqslant P = [n^{1/k}]$ pour $1 \leqslant j \leqslant s$ pour toute solution de (18) ; il n'y a donc pas lieu de faire intervenir les exposants $m^k > n$ dans la série $f(z)$, ce qui conduit à remplacer f par le *polynôme* :

$$\sum_{m=1}^{P} z^{m^k}$$

et permet, dans (19), de prendre pour C le cercle $|z| = 1$ lui-même. D'autre part, il n'y a pas intérêt à tenir compte des valeurs $x_j = 0$ dans les solutions de (18) ; finalement, si l'on note $r(n)$ le nombre de solutions de (18) pour lesquelles $x_j \geqslant 1$ pour tout j, on obtient l'expression :

$$(20) \qquad r(n) = \int_0^1 (h(x))^s e^{-2\pi inx}\,dx,$$

où :

$$h(x) = \sum_{m=1}^{P} e^{2\pi i m^k x}.$$

Pour traiter cette intégrale, Vinogradov retient le principe de la méthode de Hardy-Littlewood, en divisant l'intervalle [0, 1] en « major arcs » et « minor arcs » liés aux suites de Farey ; mais il utilise, en outre, pour majorer la contribution des « minor arcs », une méthode nouvelle, beaucoup plus puissante que la méthode de H. Weyl qu'avaient appliquée à cette fin Hardy et Littlewood.

Les résultats sont les suivants :

a) La contribution des « major arcs » fait intervenir la fonction gamma ; elle est de la forme :

$$(21) \qquad \frac{\left(\Gamma\left(1+\frac{1}{k}\right)\right)^s}{\Gamma\left(\frac{s}{k}\right)}\,\mathfrak{S}(n,s)\,n^{s/k-1} + o(n^{s/k-1}),$$

où $\mathfrak{S}(n, s)$ est un nombre dépendant de n, s et k, mais restant *borné* lorsque n varie de 1 à $+\infty$, s et k restant fixes. Ce nombre, dit « série singulière de Hardy-Littlewood », s'obtient de la façon suivante. Pour tout entier $q \geqslant 1$, on pose :

$$S(a, q) = \sum_{m=0}^{q} \exp(2\pi i m^k a/q),$$

pour tout entier a tel que $1 \leqslant a \leqslant q - 1$, premier à q ; on pose ensuite :

$$A(q) = q^{-s} \sum_{(a,q)=1} (S(a,q))^s \exp(-2\pi i n a/q).$$

la somme s'étendant aux entiers a précédents, et enfin :

$$\mathfrak{S}(n,s) = \sum_{q=1}^{\infty} A(q) ;$$

on prouve que la série est absolument convergente pour $s \geqslant 2k + 1$ et que, pour $s \geqslant 4k$, il y a un nombre $c > 0$ tel que $Q(n, s) \geqslant c$ pour tout entier n.

b) Hardy et Littlewood prouvent que, pour $s \geqslant (k-2)2^{k-1} + 5$, la contribution des « minor arcs » est $o(n^{s/k-1})$, et Vinogradov obtient le même résultat en supposant seulement que $k \geqslant 12$ et $s \geqslant 10\, k^2 \ln k$. Si l'on désigne par $G(k)$ le plus petit nombre tel que, pour $s \geqslant G(k)$, l'équation (18) ait des solutions dès que n est supérieur à un entier *n,(k)* dépendant de k, on voit donc que, pour $k \geqslant 12$, on a $G(k) \leqslant 10k^2 \ln k$, et la partie principale de $r(n)$, lorsque n tend vers $+\infty$, est donnée par le premier terme de (21) dès que $s \geqslant 10\, k^2 \ln k$. En fait, par un raffinement de ses méthodes, Vinogradov a pu montrer que l'on a l'inégalité $G(k) \leqslant k\,(3 \ln k + 11)$, sans peut-être, alors, que la partie principale de $r(n)$ soit donnée par (21).

c) On montre aisément que $G(k) \geqslant k + 1$ et on conjecture que $G(k)/k$ est borné. On a $G(2) = 4$ par le théorème de Lagrange, les nombres de la forme $8m + 7$ ne pouvant être somme de trois carrés ; Davenport a prouvé que $G(4) = 16$. Pour les autres valeurs de k, on n'a que des majorations pour G, obtenues par des procédés particuliers : $G(3) \leqslant 7$, $G(5) \leqslant 23$.

Le nombre $g(k)$ intervenant dans la formulation originale du problème de Waring a pu, grâce aux travaux de Vinogradov, être complètement déterminé, sauf pour $k = 4$ et $k = 5$; il est toujours au moins égal à $2^k - 2 + [(3/2)^k]$ et est égal à ce nombre, sauf pour un nombre fini d'exposants k. On a :

$$g(3) = 9, \quad 19 \leqslant g(4) \leqslant 35,$$
$$37 \leqslant g(5) \leqslant 54, \quad g(6) = 73.$$

Le problème de Goldbach

Sans doute sur la base d'essais numériques, un contemporain d'Euler, C. Goldbach, avait émis en 1742 la conjecture que tout entier pair est somme de deux nombres premiers et tout entier impair somme de trois nombres premiers. Aucune de ces deux conjectures n'est encore complètement démontrée, mais Vinogradov a pu établir en 1937 que tout nombre impair *assez grand* est somme de trois nombres premiers. Si l'on pose :

$$(22) \qquad f(x, n) = \sum_{p \leqslant n} \exp(2\pi i p x),$$

où la somme est étendue à tous les nombres premiers $p \leqslant n$, le nombre de solutions de l'équation $p_1 + p_2 + p_3 = n$ en nombres premiers est donné par :

$$(23) \quad r(n) = \int_{x_0}^{x_0+1} (f(x,n))^3 \exp(-2\pi i n x)\, dx,$$

pour un x_0 réel quelconque. L'évaluation de cette intégrale se fait encore en suivant l'idée de la méthode de Hardy-Littlewood ; on prend $x_0 = n^{-1} \ln^{15} n$ et, dans l'intervalle $[x_0, x_0 + 1]$, les « major arcs » sont les intervalles centrés aux points h/q (où $q \leqslant \ln^{15} n$, $0 < h < q$, h premier à q) et de demi-longueur x_0. Si E, est le complémentaire de la réunion de ces intervalles, Vinogradov commence par prouver, par

une ingénieuse et longue succession de majorations de nature élémentaire (n'utilisant même pas le théorème des nombres premiers), que l'on a :

$$(24)\quad \left|\int_{E_1} (f(x,n))^3 \exp(-2\pi i n x)\,dx\right| \leqslant Cn^2 \ln^{-4} n,$$

où C est une constante. La partie profonde du raisonnement est l'évaluation de $f(x, m)$ pour $m \leqslant n$ et pour x dans un « major arc », c'est-à-dire $x = (h/q) + y$ où q et h sont comme ci-dessus et où $|y| \leqslant x_0$. L'idée essentielle est de remarquer que, dans la somme :

$$f\left(\frac{h}{q}, m\right) = \sum_{p \leqslant m} \exp(2\pi i p h/q),$$

si l'on ne considère que les nombres premiers ne divisant pas q, l'erreur commise en valeur absolue est au plus q. Mais la somme restante s'écrit alors :

$$\sum_{\substack{0<l<q\\(l,q)=1}}^{\infty} \exp(2\pi i l h/q)\,\pi(m; q, l),$$

où $\pi(m ; q, l)$ est le nombre de nombres premiers $p \leqslant m$ qui appartiennent à la *progression arithmétique* des nombres $kq + l$ (k entier arbitraire). En utilisant (cf. *Le théorème de lu progression arithmétique*, in chap. 2) la forme la plus précise connue de l'estimation asymptotique de $\pi(m ; q, l)$, on parvient alors à l'inégalité :

$$(25)\quad \left|f\left(\frac{h}{q}+y, n\right) - \frac{\mu(q)}{\varphi(q)} g(y,n)\right| \leqslant n \ln^{-69} n,$$

où μ est la fonction de Mobius, φ la fonction d'Euler et où :

$$(26)\quad g(x, m) = \sum_{k=2}^{m} \frac{1}{\ln k} \exp(2\pi i k x).$$

Autrement dit, on a remplacé la sommation sur les *nombres premiers* $\leqslant m$ par une sommation sur tous les entiers $\leqslant m$, beaucoup plus maniable. En particulier, si l'on pose :

$$\rho(n) = \int_{-1/2}^{1/2} (g(y,n))^3 \exp(-2\pi i n y)\,dy,$$

on montre élémentairement que l'on a :

$$(27)\quad \frac{1}{2} n^2 \ln^{-3} n \leqslant \rho(n) \leqslant n^2;$$

d'autre part, on introduit, pour tout entier q, la somme :

$$(28)\quad c_q(n) = \sum_{0<h<q, (h,q)=1} \exp(-2\pi i n h/q);$$

on montre que la série :

$$(29)\quad S(n) = \sum_{q=1}^{\infty} \frac{\mu(q)}{\varphi^3(q)} c_q(n)$$

est absolument convergente et que l'on a :

$$(30)\quad S(n) = \prod_p \left(1 - \frac{c_p(n)}{(p-1)^3}\right),$$

où le produit est étendu à tous les nombres premiers ; cela donne $S(n) = 0$ pour n pair, et on montre aisément que $S(n) \geqslant 1$ pour n impair. On déduit alors de (23). (24) et (25) l'inégalité :

$$|r(n) - S(n)\rho(n)| \leqslant C' n^2 \ln^{-4} n,$$

ce qui, joint à (27), prouve que $r(n) \geqslant 1$ pour n assez grand.

2. La théorie multiplicative

Le point de vue formel

On a vu *supra* (cf. *Le point de vue formel*, in chap. 1) que le monoïde multiplicatif N * vérifie la condition (D), et qu'on peut donc définir son *algèbre large* sur un corps K ; on se bornera encore au cas où K = C,

et on notera D cette algèbre large. On note ici $n^{-\omega}$ l'élément u_n de la base canonique de C[N*], et cette fois, un élément $f \in$ D se note :

$$\sum_{n=1}^{\infty} f(n) n^{-\omega},$$

et on dit que c'est une *série formelle de Dirichlet*. Le produit de deux élémentsfet g de D se note aussi $f * g$ et est défini par :

$$(31) \qquad (f * g)(n) = \sum_{qr=n} f(q) g(r),$$

ce qui s'écrit encore :

$$(32) \quad \left(\sum_{n=1}^{\infty} f(n) n^{-\omega}\right)\left(\sum_{n=1}^{\infty} g(n) n^{-\omega}\right) = \sum_{n=1}^{\infty} \left(\sum_{qr=n} f(q) g(r)\right) n^{-\omega}$$

Pour qu'une série formelle de Dirichlet :

$$\sum_{n=1}^{\infty} f(n) n^{-\omega}$$

ait un *inverse* dans D, il faut et il suffit que $f(1) \neq 0$. *L'ordre* d'une série formelle de Dirichlet :

$$\sum_{n=1}^{\infty} f(n) n^{-\omega}$$

non nulle est encore défini comme le plus petit entier n tel que $f(n) \neq 0$. On voit qu'on peut encore définir dans D la somme *infinie (4)* et le *produit infini (5)* lorsque l'ordre de f_n tend vers $+\infty$ avec n. Si on convient d'écrire F(o) une série formelle de Dirichlet :

$$\sum_{n=1}^{\infty} f(n) n^{-\omega},$$

on écrit alors F (ω + a), pour tout nombre complexe *a*, la série formelle de Dirichlet :

$$\sum_{n=1}^{\infty} (f(n) n^{-\alpha}) n^{-\omega};$$

de même, on écrit F (ko), pour tout entier *k,* la série formelle de Dirichlet :

$$\sum_{n=1}^{\infty} (f(n^k) n^{-\omega}$$

et F'(o) la série formelle de Dirichlet :

$$\sum_{n=1}^{\infty} (-f(n) \ln n) n^{-\omega}.$$

La décomposition d'un entier en facteurs premiers exprime encore que le monoïde multiplicatif N* est « produit direct » d'une infinité de monoïdes isomorphes au monoïde additif N. Dans la théorie des séries formelles de Dirichlet, ce fait apparaît de la manière suivante. Rappelons qu'une fonctionfdéfinie dans **N***, à valeurs complexes, est dite *multiplicative* si l'on a $f(1) = 1$ et $f(mn) = f(m) f(n)$ pour deux entiers m et n premiers entre eux (cf. DIVISIBILITÉ). On vérifie aisément que, si f et g sont multiplicatives, il en est de même de $f * g$ et de l'inverse defdans D. Lorsque f est multiplicative, on a la décomposition en produit infini (produit étendu à tous les nombres premiers p) de la série formelle de Dirichlet :

$$(33) \quad F(o) = \sum_{n=1}^{\infty} f(n) n^{-\omega}$$

$$= \prod_p (1 + f(p) p^{-\omega} + \ldots f(p^k)(p^k)^{-\omega} + \ldots)$$

qui équivaut à la décomposition $f(n) = f(p_1^{k_1}) \ldots f(p_r^{k_r})$ pour la décomposition de n en facteurs premiers $n = p_1^{k_1} \ldots p_r^{k_r}$.

Rappelons (cf. DIVISIBILITÉ) que la fonction constante $i : n \mapsto 1$, la fonction de Mobius μ, la fonction d'Euler φ, la fonction $n \mapsto d(n)$, où $d(n)$ est le nombre de diviseurs de n, et les fonctions $n \mapsto$ o,(n), où o,(n) est la somme des puissances cc-ièmes des diviseurs de n, sont multiplicatives ; il en est de même de la fonction $n \mapsto 2^{\nu(n)}$, où v(n) est le nombre des facteurs premiers distincts de n, de la fonction de Liouville $n \mapsto$ h(n), définie par $\lambda(n) = (-1)^k$, où k est le nombre des facteurs premiers de n, comptés avec leur ordre de multiplicité ; enfin, la fonction de von Mangoldt $n \mapsto$ A(n) est aussi multiplicative, A(n) étant 0 si n n'est pas une puissance d'un nombre premier, égal à $\ln p$ si $n = p^m$ est une telle puissance.

À ces diverses fonctions multiplicatives correspondent des séries formelles de Dirichlet, en premier lieu la série *zêta* :

$$(34) \quad \zeta(\omega) = \sum_{n=1}^{\infty} n^{-\omega} = \prod_p (1 - p^{-\omega})^{-1},$$

et on montre que les séries formelles de Dirichlet correspondant aux autres fonctions multiplicatives s'expriment à l'aide de la série zêta par :

$$(35) \quad \sum_n \mu(n) n^{-\omega} = 1/\zeta(\omega),$$

$$(36) \quad \sum_n \mu(n)\, n^{-\omega} = \zeta(\omega)/\zeta(2\omega),$$

$$(37) \quad \sum_n \phi(n) n^{-\omega} = \zeta(\omega - 1)/\zeta(\omega),$$

$$(38) \quad \sum_n d(n) n^{-\omega} = \zeta(\omega)^2,$$

$$(39) \quad \sum_n \sigma_\alpha(n) n^{-\omega} = \zeta(\omega)\zeta(\omega - a),$$

$$(40) \quad \sum_n 2^{\nu(n)} n^{-\omega} = \zeta(\omega)^2/\zeta(2\omega),$$

$$(41) \quad \sum_n \lambda(n) n^{-\omega} = \zeta(2\omega)/\zeta(\omega),$$

$$(42) \quad \sum_n \Lambda(n) n^{-\omega} = -\zeta'(\omega)/\zeta(\omega).$$

Séries de Dirichlet

L'idée fondamentale de la théorie multiplicative est tout à fait analogue à celle de la théorie additive : si, dans une série formelle de Dirichlet :

$$\sum_n a(n) n^{-\omega},$$

on remplace le symbole $n^{-\omega}$ par le nombre complexe $n^{-s} = \exp(-s \ln n)$, où $s \in$ C, on obtient une série de nombres complexes, qui, si elle converge, a pour somme une fonction de s ; l'application de la théorie de Cauchy à cette fonction dans les régions du plan C où cette fonction est analytique donnera des informations sur les coefficients u(n) de la série.

On poses $= \sigma + it$, où σ et t sont réels, pour tout nombre complexe s, de sorte que l'on a :

$$|a(n) n^{-s}| = |a(n)|\, n^{-\sigma},$$

et, par suite, si la série de Dirichlet :

$$\sum a(n) n^{-s}$$

converge *absolument* pour $s_1 = \sigma_1 + it_1$, elle converge absolument dans le demi-plan $\sigma \geqslant \sigma_1$; on en déduit aussitôt qu'il existe un nombre réel σ_a (éventuellement égal à $+\infty$ ou $-\infty$) tel que la série converge absolument pour $\sigma > \sigma_a$ et ne converge pas absolument pour $\sigma < \sigma_a$. Mais, contrairement à ce qui se passe pour les séries entières, il se peut qu'une série de Dirichlet soit convergente sans l'être absolument dans toute une partie ouverte non vide du plan. De façon précise, on montre, par un raisonnement d'intégration par parties, que, si la série converge pour $s_0 = \sigma_0 + it_0$, elle converge *uniformément* dans tout angle de sommet s_0, défini pars $= s_0 + \rho e^{i\theta}$, avec $\rho \geqslant 0$ et $\alpha \leqslant \theta \leqslant \alpha$, α étant un nombre quelconque tel que $0 < \alpha < \pi/2$. De là on déduit aisément qu'il existe un nombre réel $\sigma_c \leqslant \sigma_a$ tel que la série converge pour $\sigma > \sigma_c$ et ne converge pas pour $\sigma < \sigma_c$; en

outre, on a toujours $\sigma_a \quad \sigma_c \leqslant 1$. La fonction :

$$F(s) = \sum_{n=1}^{\infty} a(n) n^{-s}$$

est alors *holomorphe* pour $\sigma > \sigma_c$, et ses dérivées s'obtiennent en dérivant la série de Dirichlet terme à terme dans ce demi-plan. Lorsque la fonction a(n) est multiplicative, la formule :

$$(43)\quad F(s) = \prod (1 + a(p)p^{-s} + + a(p^k)p^{-ks} + \ldots)$$

est valable pour $\sigma > \sigma_a$, le produit infini du second membre (dit « *produit eulérien* » de F) étant uniformément convergent dans tout demi-plan $\sigma \leqslant \sigma_a + \varepsilon$ pour $\varepsilon > 0$.

La théorie de Cauchy ne fournit pas ici directement les coefficients $a(n)$ à l'aide d'une intégrale curviligne, mais seulement les *sommes* de ces coefficients :

$$(44)\qquad A(m) = \sum_{n \leqslant m} a(n).$$

On montre en effet que, pour tout nombre a > 0 tel que $a > \sigma_a$ et tout m entier $\leqslant 1$, on a, pour tout $T > 0$,

$$(45)\quad \frac{1}{2\pi i}\int_{\alpha - iT}^{\alpha + iT} \frac{(m + 1/2)^s}{s} F(s)\, ds - A(m) \leqslant \frac{(m + 1)^{\alpha + 1}}{T} \sum_{n=1}^{\infty} \frac{|a(n)|}{n^{\alpha}}.$$

Le théorème des nombres premiers

À la fin du XVIIIe siècle, A. M. Le Gendre et C. F. Gauss, indépendamment, avaient émis la conjecture (d'après les tables de nombres premiers) que le nombre $\pi(x)$ des nombres premiers $\leqslant x$ avait, lorsque x tend vers $+\infty$, une partie principale :

$$(46)\qquad \pi(x) \sim \frac{x}{\ln x};$$

Gauss avait même précisé cette conjecture en indiquant comme meilleure approximation de $\pi(x)$ la fonction li, appelée logarithme intégral :

$$\mathrm{li}(x) = \int_2^x \frac{dt}{\ln t}.$$

Le théorème des nombres premiers établit (46) ; démontré d'abord en 1896 par J. Hadamard et C. de La Vallée-Poussin indépendamment, il a été par la suite amélioré par divers mathématiciens et l'on peut maintenant montrer que :

$$(47)\quad \pi(m) = \mathrm{li}(m) + O(m \exp(-c\sqrt{\ln m})),$$

où $c > 0$ est une constante. Si l'hypothèse de Riemann était vraie (cf. fonction ZÊTA), on pourrait remplacer dans (47) le terme complémentaire par $O(m^{1/2} \ln m)$: on sait que c'est la meilleure majoration possible.

La démonstration de (47) utilise la formule (45) appliquée à la série de Dirichlet :

$$F(s) = -\zeta'(s)/\zeta(s) = \sum_{n=1}^{\infty} A(n) n^{-s}.$$

Si l'on pose :

$$\theta(m) = \sum_{p \leqslant m} \ln p,$$

où p parcourt les nombres premiers $\leqslant m$, et $A(n) = \theta(n) - n$, on a :

$$\pi(m) = \sum_{n=2}^{m} \frac{\theta(n) - \theta(n-1)}{\ln n}$$

$$= \sum_{n=2}^{n} \frac{1 + A(n) - A(n-1)}{\ln n}$$

$$= \left(\frac{1}{\ln 2} + \frac{1}{\ln 3} + \ldots + \frac{1}{\ln m}\right) + \frac{\Delta(m)}{\ln m} - \frac{\Delta(1)}{\ln 2} + \sum_{n=2}^{m-1} \Delta(n)\left(\frac{1}{\ln n} - \frac{1}{\ln(n+1)}\right).$$

Dans cette dernière expression, la première somme diffère de li (m) par une quantité bornée. Pour prouver (47), il suffit donc d'établir que :

$$|\Delta(n)| \leqslant Cn \exp(-c\sqrt{\ln n}),$$

pour deux constantes C > 0 et c > 0.

Or, si l'on pose :

$$U'(m) = \sum_{n \leqslant m} \Lambda(n),$$

on montre élémentairement que :

$$0 \leqslant \Psi(m) - \theta(m) \leqslant m^{1/2} \ln^2 m$$

et on est donc ramené à montrer que l'on a :

$$U'(m)-m \leqslant cm \exp(-c\sqrt{\ln m}).$$

On applique (45) en prenant :

$$\alpha = 1 + \frac{1}{\ln\left(m + \frac{1}{2}\right)},$$

$$T = \exp\left(\frac{1}{100}\sqrt{\ln\left(m + \frac{1}{2}\right)}\right).$$

On sait que la fonction ζ se prolonge en une fonction méromorphe dans C, ayant un seul pôle simple au point $s = 1$ et ne s'annulant pas dans le demi-plan $\sigma > 1$ (cf. fonction ZÊTA) ; la série de Dirichlet F(s) converge absolument dans ce demi-plan, et l'on montre aisément à partir de (45) que l'on a, en posant $\eta(s) = (s-1)\zeta(s)$,

$$\Psi(m) - m + \frac{1}{2\pi i}\int_{\alpha - iT}^{\alpha + iT} \frac{(m+1/2)^s}{s}\frac{\eta'(s)}{\eta(s)}ds \leqslant C'm \exp\left(-\frac{1}{200}\sqrt{\ln m}\right).$$

Le point essentiel est de majorer l'intégrale qui subsiste dans cette formule. La propriété de la fonction $\eta(s)$ que l'on utilise pour cela est la suivante :

Dans l'ensemble D (cf. figure, où les proportions ne sont pas respectées), défini par les inégalités :

$$1 - \frac{1}{10\,000 \ln t^*} \leqslant \sigma \leqslant 2,$$

où $t^* = \max(|t|, 100)$, la fonction η ne s'annule pas, et l'on a l'inégalité :

$$(48) \qquad \left|\frac{\eta'(s)}{\eta(s)}\right| \leqslant C_1 \ln^3 t^*,$$

ou C_1 est une constante.

Cela s'établit à l'aide de raisonnements assez longs, mais élémentaires, de la théorie des fonctions holomorphes d'une variable complexe, à partir des inégalités suivantes :

$$(49) \qquad \left|\zeta(s) - \frac{1}{s-1}\right| \leqslant \frac{|s|}{\sigma}, \quad \sigma > 0,\ s \neq 1,$$

$$(50) \qquad |\zeta^3(\sigma)\zeta^4(\sigma + it)\zeta(\sigma + 2it)| \geqslant 1,\ \sigma > 1,$$

la seconde inégalité étant conséquence de l'observation que la partie réelle de $3 + 4e^{i\theta} + e^{2i\theta}$ est toujours $\leqslant 0$ pour θ réel.

Cela étant, on utilise (48) de la façon suivante. En vertu du théorème de Cauchy appliqué à la fonction :

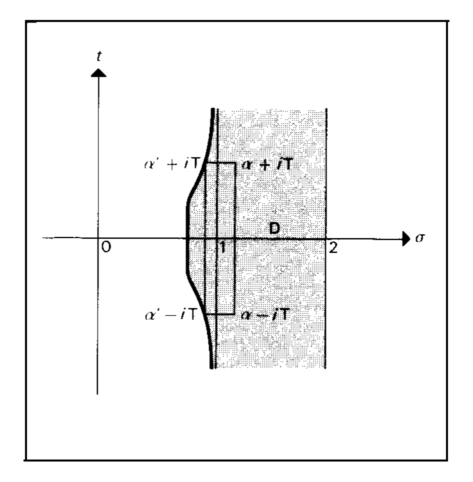
$$\frac{(m+1/2)^s}{s}\frac{\eta'(s)}{\eta(s)}$$

holomorphe dans un voisinage ouvert de D, l'intégrale de cette fonction, étendue au segment d'extrémités $\alpha' - iT$, $\alpha' + iT$, est égale à l'intégrale étendue aux trois autres côtés du rectangle d'extrémités $\alpha \pm iT$, $\alpha' \pm iT$, pourvu que ce rectangle soit contenu dans D (cf. figure). On prend :

$$\alpha' = 1 - \frac{1}{10\,000 \ln T}$$

qui répond à la question pour m assez grand. Il est alors facile, en appliquant la majoration (48), de voir que l'on a bien :

$$\left|\int_{\alpha - iT}^{\alpha + iT} \frac{(m+1/2)^s}{s}\frac{\eta'(s)}{\eta(s)}ds\right| \leqslant C_2 m \exp\left(-\frac{1}{200}\sqrt{\ln m}\right),$$



pour une constante convenable C_2, ce qui achève la démonstration.

Le théorème de la progression arithmétique

La méthode d'Euclide prouvant l'existence d'une infinité de nombres premiers peut, convenablement modifiée, établir par exemple qu'il y a une infinité de nombres premiers de la forme $4n + 3$ ou de la forme $6n + 5$. Le théorème de la progression arithmétique affirme que, quels que soient les entiers k et l premiers entre eux, il y a une infinité de nombres premiers de la forme $kn + l$; il fut démontré pour la première fois en 1837 par Dirichlet, qui, à cette occasion, introduisit à la fois dans la science les deux notions de série de Dirichlet et de *caractère* d'un groupe abélien fini (cf. GROUPES - Représentation linéaire des groupes) ; les améliorations et généralisations de ce théorème utilisent toujours l'idée extrêmement originale de Dirichlet.

La forme la plus précise du théorème de Dirichlet est la suivante :

Notons $\pi(m\ ;\ k,\ l)$ le nombre de nombres premiers $p \leqslant m$ qui sont de la forme $kn + l$; le nombre $\pi(m\ ;\ 1,\ 1)$ n'est autre que le nombre $\pi(m)$ introduit plus haut. Alors, pour tout entier N, on peut écrire :

$$(51)\quad \left|\pi(m\,;k,l)-\frac{1}{\varphi(k)}\,\mathrm{li}\,(m)\right| \leqslant C(N)m\exp\left(-\frac{1}{200}\sqrt{\ln m}\right),$$

où la constante C(N) ne dépend que de N, l'inégalité étant valable pour tout entier m, tout entier k tel que :

$$(52)\qquad k \leqslant \ln^{N} m$$

et tout entier $1 < k$ premier avec k ; φ est la fonction d'Euler, <p(k) étant donc le nombre d'entiers l tels que $1 \leqslant l < k$, qui sont premiers avec k. La comparaison des parties principales de $\pi(m\ ;\ k,\ l)$ et de $\pi(m)$ montre que l'on peut encore dire que les nombres premiers sont « *également répartis* » dans les $q(k)$ progressions arithmétiques $kn + l$.

L'idée fondamentale de Dirichlet est de considérer à lu *fois* ces <p(k) classes d'entiers et d'utiliser le fait qu'elles forment naturellement un *groupe commutatif,* savoir le groupe $(\mathbf{Z}/k\mathbf{Z})^*$ des éléments inversibles de l'anneau $\mathbf{Z}/k\mathbf{Z}$. 11 y a donc *<p(k) caractères* χ_h, *1* $\leqslant h \leqslant$ *<p(k),* de ce groupe (à valeurs dans le groupe U des nombres complexes de valeur absolue 1), qui ici sont des *homomorphismes* de $(\mathbf{Z}/k\mathbf{Z})^*$ dans U (ce qui entraîne que leurs valeurs sont des racines $\varphi(k)$*-ièmes de l'unité*) ; rappelons que l'on a les *relations d'orthogonalité* :

$$(53)\quad \sum_{h\leqslant 1}^{\varphi(k)} \overline{\chi_h(x)}\,\chi_h(y) = \begin{cases} 0, & x \neq y, \\ \varphi(k), & x = y, \end{cases}$$

pour x, y dans $(\mathbf{Z}/k\mathbf{Z})^*$,

$$(54)\quad \sum_{x} \overline{\chi_j(x)}\,\chi_h(x) = \begin{cases} 0, & j \neq h, \\ \varphi(k), & j = h, \end{cases}$$

la somme étant étendue à tous les éléments x de $(\mathbf{Z}/k\mathbf{Z})^*$. L'application $\chi_1 : x \longrightarrow 1$ est

un caractère dit *principal* et, pour tout caractère *non principal* χ, on a, d'après (54),

$$\sum \chi(x) = 0.$$

À partir de ces caractères de $(\mathbf{Z}/k\mathbf{Z})^*$, on définit *sur* $\mathbf{Z}$ des fonctions $\chi(n)$ dits « caractères modulo k » en prenant $\chi(n)$ égal à $\chi(x)$ où x est la classe de n (mod k) si $(n, k) = 1$, et $\chi(n) = 0$ si $(n. k) \neq 1$. On vérifie aussitôt que $\chi(mn) = \chi(m)\chi(n)$ quels que soient les entiers m et n ; le « caractère principal modulo k » est la fonction égale à 1 pour $(n, k) = 1$, à 0 sinon.

Cela étant, on peut écrire :

$$\pi(m\,;\,k,\,l) = \sum_{p \leqslant m} g(p),$$

où p parcourt l'ensemble des nombres premiers $\leqslant m$ et $g(n) = 1$ si $n \equiv l$ (mod k), $g(n) = 0$ sinon. Or. en raison des relations d'orthogonalité (53), on peut écrire :

$$(55) \qquad g(n) = \frac{1}{\varphi(k)} \sum_{h=1}^{\phi(k)} \overline{\chi_h(l)}\,\chi_h(n),$$

d'où :

$$(56) \quad \pi(m\,;k,l) = \frac{1}{\varphi(k)} \sum_{h=1}^{\phi(k)} \overline{\chi_h(l)}\,\pi(m\,;\chi_h),$$

où l'on a posé :

$$(57) \qquad \pi(m\,;\chi_h) = \sum_{p \leqslant m} \chi_h(p).$$

Si χ_1 est le caractère principal modulo k, on a d'autre part :

$$(58) \qquad |\,\pi(m) - \pi(m\,;\chi_1)\,| \leqslant \varphi(k),$$

pour $k \leqslant m$; en vertu du théorème des nombres premiers sous la forme (47), on voit que, pour établir (51), il suffit de montrer que l'on a :

$$|\,\pi(m\,;\chi_h)\,| \leqslant C_1(N)m \exp\left(-\frac{1}{200}\sqrt{\ln m}\right),$$

pour $h = 2,3, \ldots, q(k)$ et $k \leqslant \ln^N m$.

La marche suivie est analogue à celle du théorème des nombres premiers, en remplaçant la fonction $\zeta(s)$ par les « fonctions L » de Dirichlet :

$$(59) \quad L(s,\,\chi) = \sum_{n=1}^{\infty} \chi(n)n^{-s} = \prod (1 - \chi(p)p^{-s})^{-1},$$

définies pour chaque k et chaque « caractère modulo k » χ. Il suffit de se borner aux caractères non principaux ; on montre alors aisément que la série de Dirichlet (59) converge dans le demi-plan $\sigma > 0$ (contrairement à ce qui se passe pour un caractère principal) et que l'on a dans ce demi-plan l'inégalité analogue à (49) :

$$(60) \qquad |\,L(s,\chi)\,| \leqslant k\,|\,s\,|\,/\sigma,$$

si χ est un caractère modulo k.

La difficulté est d'obtenir une inégalité analogue à (48) pour la dérivée logarithmique $L'(s,\,\chi)/L(s,\,\chi)$ dans un domaine D_N analogue au domaine D considéré dans la figure *supra,* mais dépendant de N ainsi que la constante C_1, l'inégalité devant être satisfaite pour *tous* les caractères non principaux modulo k et *tous* les entiers k vérifiant (52). Le point capital est de montrer que les $L(s,\,\chi)$ ne s'annulent pas dans un tel domaine D_N ; on y parvient assez facilement pour les points d'un tel domaine non situés sur l'axe réel en utilisant des inégalités analogues à (50) pour les fonctions L. On prouve aussi (comme l'avait déjà fait Dirichlet) que $L(1,\chi) \neq 0$ pour tout caractère non principal, ce qui permet d'obtenir (5 1) pour un *k fixé*. Pour avoir la forme générale

de (51), il faut faire appel à un résultat plus profond, démontré par C. L. Siegel en 1936 : Pour tout $\varepsilon > 0$, il y a un entier $A(\varepsilon)$ tel que, pour *tout* $k \geqslant A(\varepsilon)$ et *tout* caractère non principal χ modulo k, on ait $L(s, \chi) \neq 0$ pour $1 - k^{-\varepsilon} \leqslant s \leqslant 1$.

Mentionnons ici un résultat démontré par Y. Linnik au moyen d'autres méthodes : *le plus petit* nombre premier appartenant à une progression arithmétique $kn + l$, pour k et l premiers entre eux, est majoré par k^c, où c est une constante.

Le nombre de classes d'idéaux d'un corps quadratique

Les résultats précédents peuvent se généraliser aux idéaux premiers d'un corps de nombres algébriques, grâce à l'extension à ces corps des définitions de la fonction zêta et des fonctions L. Nous ne mentionnerons ici qu'un cas particulier des résultats de cette théorie, le lien découvert par Dirichlet entre les fonctions L et le nombre de classes d'idéaux d'un corps quadratique. De façon précise, le discriminant d d'un corps quadratique est non divisible par un carré $\neq 4$, et est ou multiple de 4, ou de la forme $4k + 1$; pour tout entier impair $n > 1$, on désigne par $\left(\frac{d}{n}\right)$ le symbole de Jacobi (cf. DIVISIBILITÉ) si d et n sont premiers entre eux, 0 sinon ; on définit $\left(\frac{d}{2}\right)$ comme égal à 1 si $d \equiv 1 \pmod 8$, à -1 si $d \equiv 1 \pmod 8$, et à 0 dans les autres cas ; on montre alors que $n \mapsto \left(\frac{d}{n}\right)$ est un caractère non principal modulo $|d|$, et on peut donc former la série correspondante :

$$L_d(s) = \sum_{n=1}^{\infty} \left(\frac{d}{n}\right) n^{-s},$$

et la formule découverte par Dirichlet pour le nombre de classes $h(d)$ du corps quadratique de discriminant d est :

$$(61) \qquad L_d(1) = \alpha h(d),$$

où $\alpha = \pi/\sqrt{|d|}$ si $d < -4$, $\alpha = \pi/3\sqrt{3}$ pour $d = 3$, $\alpha = \pi/4$ pour $d = -4$ et $\alpha = 2 \ln \varepsilon_d/\sqrt{d}$ si $d > 0$, ε_d étant l'« unité fondamentale » du corps. C'est à l'aide de cette formule que Dirichlet prouva que $L(1, \chi) \neq 0$ pour tout caractère non principal.

L'application la plus intéressante de la formule (61) a été donnée par Siegel ; une des formes de son théorème cité plus haut implique que :

$$L_d(1) = o(\ln |d|),$$

lorsque d tend vers $+\infty$; on en déduit. par (61),

$$(62) \qquad \ln h(d) \sim 1/2 \ln |d|,$$

lorsque d tend vers $-\infty$, et :

$$(63) \qquad \ln(h(d) \ln \varepsilon_d) \sim 1/2 \ln d,$$

lorsque d tend vers $+\infty$.

L'équivalence (62) prouve entre autres une conjecture de Gauss, selon laquelle h(d) croît indéfiniment lorsque d tend vers $-\infty$ (conjecture démontrée d'abord par H. Heilbronn). En particulier, il n'y a qu'un nombre fini de corps quadratiques de discriminant $d < 0$ dont le nombre de classes d'idéaux est donné ; pour $h(d) = 1$, on sait, d'après Stark et Baker (1966), que les seuls corps quadratiques correspondent aux valeurs $-3, -4, -7, -8, -11, -19$, 43, 67 et 163 de d.

3. Valeurs moyennes de fonctions arithmétiques

L'irrégularité des fonctions arithmétiques

Les fonctions définies dans l'ensemble des entiers > 0 par des conditions de nature arithmétique, telieslesfonctions multiplicatives qu'on a étudiées plus haut (cf. chap. 2, *Le point de vue formel),* ont une allure en

général très irrégulière. Par exemple, la fonction d(n) est égale à 2 pour n premier, mais elle est très grande pour les nombres de la forme m ! ; on peut montrer par des procédés élémentaires (n'utilisant pas le théorème des nombres premiers) que l'on a :

$$\limsup_{n\to\infty} \frac{\ln d(n) \ln \ln n}{\ln n} = \ln 2. \tag{64}$$

Pour $\sigma_1(n)$, l'irrégularité est moins prononcée ; on a o,(n) $= n + 1$ si n est premier, et on montre (à l'aide du théorème des nombres premiers) que :

$$\limsup_{n\to\infty} \frac{\sigma_1(n)}{\ln \ln n} = e^{\gamma}, \tag{65}$$

où y est la constante d'Euler (cf. fonction GAMMA). De même, pour la fonction d'Euler <p(n), on a $\varphi(n) = n(1 - 1/p)$ pour $n = p^k$, puissance d'un nombre premier, ce qui entraîne :

$$\limsup_{n\to\infty} \frac{\varphi(n)}{n} = 1;$$

on montre ici que l'on a :

$$\liminf_{n\to\infty} \frac{\varphi(n) \ln \ln n}{n} = e^{-\gamma} \tag{66}$$

Une fonction arithmétique très étudiée, mais sur laquelle on sait encore peu de chose, est la différence $p_{n+1} - p_n$ entre deux nombres premiers consécutifs, On conjecture qu'il y a une infinité de valeurs de n pour lesquelles $p_{n+1} - p_n = 2$ (nombres premiers ((jumeaux »)) et que le nombre des $p_n \leqslant$ x ayant cette propriété est asymptotiquement égal à $Cx/(\ln x)^2$, avec :

$$C = 4 \prod_p \left(1 - \frac{1}{(p-1)^2}\right)$$

(produit étendu aux nombres premiers impairs) ; mais tout ce que l'on a pu prouver jusqu'ici, avec V. Brun, est que la série des inverses des nombres premiers jumeaux est convergente. Dans l'autre direction, on peut, par une étude poussée de la fonction $\zeta(s)$, montrer que :

$$p_{n+1} - p_n = O(p_n^{3/5});$$

si l'hypothèse de Riemann était vraie, elle entraînerait $p_{n+1} - p_n = O(p_n^{1/2} \ln$ p,,).

Moyennes des fonctions arithmétiques

On espère en général que, lorsqu'une fonction f définie pour les entiers > 0 a une allure irrégulière, la fonction F(m) $= f(1) + f(2) + \ + f(m)$, égale à m fois la « valeur moyenne » de f dans l'intervalle $1 \leqslant n \leqslant m$, se comportera de façon plus satisfaisante ; c'est ce qui se passe pour la plupart des fonctions arithmétiques. Le théorème des nombres premiers en est un exemple : de façon générale, si P est une partie de N et si l'on prend pour f'la fonction caractéristique de P, la fonction correspondante $m \mapsto F(m)/m$ est ce qu'on appelle la *densité* de P dans l'intervalle $[1, m]$; le théorème des nombres premiers dit que, pour l'ensemble P des nombres premiers, cette « densité » a une partie principale $1/\ln m$.

Pour certaines fonctions considérées *supra* (cf. *Le point de vue formel*, in chap. 2), on peut effectivement obtenir des parties principales de leurs « moyennes » de façon élémentaire ; par exemple, on a :

$$\sum_{n\leqslant x} d(n) = x \ln x + (2\gamma - 1)x + O(\sqrt{x}), \tag{67}$$

$$\sum_{n\leqslant x} \sigma_1(n) = \frac{\pi^2}{12} x^2 + O(x \ln x), \tag{68}$$

$$\sum_{n\leqslant x} \varphi(n) = \frac{2}{\pi^2} x^2 + O(x \ln x), \tag{69}$$

$$\sum_{n\leqslant x} |\mu(n)| = \frac{6}{\pi^2} x + O(\sqrt{x}), \tag{70}$$

la derniére somme étant la « densité » dans l'intervalle $[1, x]$ de l'ensemble des entiers sans facteur carré.

À l'aide du théorème des nombres premiers, on établit :

(71) $$\sum_{n \leqslant x} p(n) = O(x \exp(-c\sqrt{\ln x})) = o(x),$$

(72) $$\sum_{n \leqslant x} h(n) = O(x \exp(-c\sqrt{\ln x})) = o(x),$$

(73) $$\sum_{n \leqslant x} v(n) = x \ln \ln x + Bx + o(x),$$

avec :

$$B = \gamma + \sum_{p} \left(\ln\left(1 - \frac{1}{p}\right) + \frac{1}{p}\right)$$

(somme étendue aux nombres premiers).

Le premier membre de (67) admet une interprétation géométrique intéressante : c'est le nombre des points (r, s) du *réseau* Z^2 dans le plan qui appartiennent à la région du plan formée des points (u, v) vérifiant $u \geqslant 1/2$, $v \geqslant 1/2$ et $uv \leqslant x$. Cette interprétation suggère aussitôt que la partie principale de ce nombre est l'aire de la région précédente et conduit à améliorer la majoration du terme complémentaire. D'une façon générale, considérons dans le plan la région G définie par $s \leqslant u \leqslant w$, $q \leqslant v \leqslant f(u)$, où $s - 1/2$, $w - 1/2$ et $q - 1/2$ sont entiers ; f est supposée deux fois continûment différentiable, sa dérivée étant > 0 dans $[s, w]$, et sa dérivée seconde ne s'annulant pas dans cet intervalle.

Soit I(G) le nombre de points du réseau Z^2 dans G et A(G) l'aire de G. On a alors l'inégalité suivante, établie par Van der Corput :

(74) $$A(G) - I(G) \leqslant Cr^{2/3} \sup f'(u),$$

où C est une constante absolue et r un nombre vérifiant les relations $r > 1$, $r > \sup(1/|f''(u)|)$ et $r > \sup(1/f'(u))^3$. L'idée de la démonstration consiste à faire varier le domaine G en remplaçant la fonction f(u) par $f(u) + y$, avec $y \leqslant r^{-1/3}$; soit G_y cette région variable. Si l'on pose pour simplifier $\theta = r^{-1/3}$, $\mu = \sup f'(u)$, on voit aussitôt que l'on a :

$$\left| A(G) - \frac{1}{\theta}\int_0^\theta A(G_y)\,dy \right| \leqslant r^{1/3}\mu \,;$$

on a d'autre part :

$$I(G) \leqslant \frac{1}{\theta}\int_0^\theta I(G_y)\,dy \,;$$

on majorera donc I(G) − A(G) si l'on sait majorer :

$$\int_0^\theta (I(G_y) - A(G_y))\,dy,$$

et on procédera de même pour A(G) − I(G) en donnant à y des valeurs négatives. Le point essentiel est que l'on peut écrire l'intégrale :

$$\int_0^\theta I(G_y)\,dy$$

sous forme d'une *série double :*

(75) $$\int_0^\theta I(G_y)\,dy = \sum_{m,n} \int_0^\theta P(y, m, n)\,dy,$$

où :

$$P(y, m, n) = \iint_{G_y} \cos 2\pi mu \cos 2\pi nv \,du\,dv,$$

à l'aide du développement en série de Fourier de $x - 1/2$. Or on a $P(y, 0, 0) = A(G_y)$; tout revient à majorer au second membre de (75) la somme des termes *autres* que le terme principal correspondant à $m = n = 0$. On y parvient grâce aux hypothèses faites sur f et grâce à un

lemme de Van der Corput, d'après lequel, pour une fonction réelle F, on obtient :

(76) $$\left|\int_{\alpha}^{\beta} e^{i F(t)}\, dt\right| \leqslant C/\sqrt{\lambda},$$

où C est une constante absolue, si, dans l'intervalle $[\alpha, \beta]$, on a $F''(t) \geqslant \lambda > 0$. Appliquée convenablement à la fonction $f(u) = x/u$, l'inégalité (74) permet (théorème de Voronoï) de remplacer, dans (67), $O(\sqrt{x})$ par $O(x^{1/3} \ln x)$. Si l'on prend pour G la région limitée par une ellipse $au^2 + bv^2 = x$, on obtient l'expression :

(77) $$I(G) = A(G) + O(x^{1/3}).$$

En outre, Jarnik a pu montrer que, dans l'inégalité (74), il n'est pas possible de remplacer l'exposant 2/3 par un nombre plus petit, ce qui tend à conjecturer que, dans (77), l'exposant 1/3 est le meilleur possible. Mais on a particulièrement étudié le cas $a = b = 1$; si U(n) est le nombre de solutions de l'équation $u^2 + v^2 = n$ (cf. *Sommes de* carrés, in chap. 1), I(G) est, dans ce cas, la somme :

$$\sum_{n \leqslant x} U(n) = I(x)$$

et Van der Corput a pu prouver qu'il existe un $a > 0$ tel que :

$$I(X) = \pi x + O(x^{1/3-\alpha})$$

D'autre part, Hardy a démontré que, dans cette formule, on ne peut pas remplacer $1/3 \quad a$ par 1/4 (on conjecture cependant que n'importe quel exposant > 1/4 est possible). Ces deux résultats découlent d'une analyse extrêmement subtile, à partir de la remarquable identité de Hardy :

(78) $$\frac{I(x+) + I(x-)}{2} = \pi x + \sqrt{x} \sum_{n=1}^{\infty} \frac{U(n)}{\sqrt{n}} J_1(2\pi\sqrt{nx}),$$

où J_1 désigne la fonction de Bessel d'indice 1.

Interprétations probabilistes

Étant donné une fonction réelle mesurable fdéfinie pour $0 \leqslant t < 1$, pour tout nombre ω tel que $\infty \leqslant \omega \leqslant +\infty$, on peut définir la « probabilité » pour que $f(t) < \omega$ comme la *mesure* de l'ensemble des t vérifiant cette condition : si on désigne ce nombre par a(o), la fonction σ est une fonction croissante et continue à gauche dans la droite réelle achevée $\overline{\mathbf{R}}$ (cf. espaces MÉTRIQUES, chap. 1) telle que $\sigma(-\infty) = 0$ et $\sigma(+\infty) = 1$ et est dite « fonction de répartition » de f. La « moyenne » :

$$\int_0^1 f(t)\, dt$$

de f est alors égale à :

$$\int_{-\infty}^{+\infty} x\, d\sigma(x).$$

Les « lois limites » du calcul des probabilités affirment que, sous certaines conditions, pour une suite (f_n) de fonctions mesurables, les fonctions de répartition σ_n correspondantes *convergent* vers une fonction de répartition, dont la plus connue est la fonction de Gauss :

$$G(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\omega} e^{-x^2/2}\, dx.$$

Considérons maintenant une fonction réelle h définie dans l'ensemble N des entiers $\geqslant 0$. Pour tout entier N et tout $\omega \in \overline{\mathbf{R}}$, soit $K_N(\omega)$ le nombre des entiers $n \leqslant N$ tels que $h(n) < \omega$; la fonction :

$$\sigma_N(\omega) = \frac{1}{N} K_N(\omega)$$

est la fonction de répartition de la fonction $f_N(t)$ définie, pour $0 \leqslant t < 1$, par les conditions :

$$f_N(t) = h(n),$$

pour :

$$\frac{n}{N} \leqslant t < \frac{n+1}{N}, \quad 0 \leqslant n \leqslant N-1.$$

Si ces fonctions σ_N tendent vers une limite σ lorsque N tend vers $+\infty$, cela signifie donc que, pour tout ω, la suite des entiers n telle que h(n) < ω a une *densité* $\sigma(\omega)$; on dit alors que h a une « fonction de répartition » σ.

On dit que h est *additive* si e^h est multiplicative, soit $h(mn) = h(m) + h(n)$ si m et n sont premiers entre eux. Erdös et Wintner ont montré que, pour de telles fonctions, on peut, en utilisant le théorème des nombres premiers, donner la condition nécessaire et suffisante suivante pour l'existence d'une fonction de répartition. Les deux séries :

$$(79) \qquad \sum_p \frac{h^+(p)}{p}, \quad \sum_p \frac{(h^+(p))^2}{p}$$

doivent être *convergentes* ; ici p parcourt l'ensemble des nombres premiers, avec $h^+(p)$ = h(p) si $h(p) \leqslant 1$, et $h^+(p) = 1$ dans le cas contraire. En outre, d'après P. Lévy, la fonction de répartition σ est continue si et seulement si la somme des inverses $1/p$ des nombres premiers pour lesquels $f(p) \neq 0$ est infinie.

Des exemples, où ces conditions sont vérifiées, sont donnés par les deux fonctions :

$$\ln \frac{\varphi(n)}{n}, \quad \ln \frac{\sigma_1(n)}{n};$$

pour la première de ces deux fonctions, la fonction de répartition o(o) est en outre « singulière » au sens de Lebesgue (autrement dit, Erdos a montré qu'elle a une dérivée nulle presque partout, sans être constante).

Si la seconde des séries (79) converge, mais non la première, il y a encore une fonction de répartition pour la fonction :

$$h(n) - \sum_{p \leqslant n}^{\infty} \frac{h^+(p)}{p}.$$

Enfin, si les deux séries divergent. il faut considérer la fonction :

$$g(n) = (h(n) - A(n))/B(n),$$

où l'on pose :

$$A(n) = \sum_{p \leqslant n} \frac{h(p)}{p}, \quad B(n) = \left(\sum_{p \leqslant n} \frac{h^2(p)}{p} \right);$$

il y a alors toujours. d'après Erdos et Kac, une fonction de répartition pour g égale à la fonction de Gauss G(o). Par exemple, si l'on prend pour /z(n) la fonction v(n) et si $K_N(\omega)$ est l'ensemble des entiers $n \leqslant N$ tels que :

$$v(n) \leqslant \ln \ln n + \omega \sqrt{\ln \ln n},$$

alors $K_N(\omega)/N$ tend vers $G(\omega)$, et on peut même prouver la «loi du logarithme itéré » :

$$(80) \quad \frac{K_N(\omega)}{N} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\omega} e^{-x^2/2} dx = O\left(\frac{1}{\sqrt{\ln \ln N}}\right).$$

JEAN DIEUDONNÉ

Bibliographie

Voir la bibliographie à la fin de la partie C ci-après.

B. Nombres p-adiques

On peut aborder l'étude d'un problème diophantien (cf. équations DIOPHANTIENNES) en commençant par chercher les solutions modulo p, un nombre premier **quelconque** : on est alors devant un problème plus facile, car Z/pZ est un **corps** (cf. DIVISIBILITÉ). Cette

méthode ne donne qu'une information insuffisante pour le problème initial ; on la raffine en étudiant les équations modulop"' pour tous les entiers $m \geqslant 1$. L'anneau $\mathbf{Z}/p^m\mathbf{Z}$ n'est pas un corps, mais ses propriétés arithmétiques sont beaucoup plus simples que celles de Z : c'est un anneau fini qui a un seul idéal premier (engendré par la classe dep) ; les autres idéaux sont les puissances de l'idéal premier.

Supposons maintenant qu'on connaisse une solution $x_m \in \mathbf{Z}/p^m\mathbf{Z}$ du problème modulo p^m ; pour tout $k \leqslant m$, on en déduit une solution x_k mod p^k au moyen de l'application canonique évidente :

$$\varphi_{km} : \mathbf{Z}/p^m\mathbf{Z} \rightarrow \mathbf{Z}/p^k\mathbf{Z},$$

provenant de l'application identique de Z. Ces considérations nous conduisent à introduire les suites (x_m) telles que $x_m \in \mathbf{Z}/p\ \mathbf{Z}$ pour tout m et $x_k = \varphi_{km}(x_m)$ pour $k \leqslant$ m. L'ensemble $\mathbf{Z}_p$ de ces suites est ainsi une partie du produit :

$$\prod_m \mathbf{Z}/p^m\mathbf{Z},$$

et c'est même un sous-anneau pour la structure d'anneau produit (car les φ_{km} sont des homomorphismes d'anneaux) ; on dit que c'est la limite projective des anneaux $\mathbf{Z}/p^m\mathbf{Z}$. L'anneau $\mathbf{Z}_p$ ainsi défini s'appelle l'*anneau* ***des entiers p-adiques*** ; la méthode d'approche d'un problème diophantien envisagée plus haut consiste à étudier d'abord le problème dans $\mathbf{Z}_p$. Dans cet article, on donne les principales propriétés arithmétiques de cet anneau.

1. Généralités

Le noyau de l'homomorphisme canonique $\mathbf{Z} \rightarrow \mathbf{Z}_p$ est formé des entiers divisibles par toutes les puissances dep ; il est donc réduit à {O), et l'homomorphisme considéré est injectif et permet d'identifier Z à un sous-anneau de Z,,. La surjection canonique $\mathbf{Z} \rightarrow \mathbf{Z}/p^n\mathbf{Z}$ se décompose en l'injection de Z dans $\mathbf{Z}_p$ suivie de la projection de $\mathbf{Z}_p$ dans le facteur $\mathbf{Z}/p^n\mathbf{Z}$; on voit ainsi que cette dernière projection est surjective ; son noyau est l'ensemble des entiers p-adiques (x_m) tels que $x_n = 0$, ce qui donne $x_m = 0$ pour $m \leqslant n$ et $x_m \in p^n\mathbf{Z}/p^m\mathbf{Z}$ pour $m > n$; autrement dit, ce noyau est l'ensemble des entiers p-adiques multiples de p^n. On obtient ainsi l'isomorphisme :

$$\mathbf{Z}_p/p^n\mathbf{Z}_p \simeq \mathbf{Z}/p^n\mathbf{Z}$$

pour tout entier $n \geqslant 1$.

Pour $n = 1$, cela montre que $\mathbf{Z}_p/p\mathbf{Z}_p$ **est** un corps, donc que l'idéal $p\mathbf{Z}_p$ engendré par p est maximal. Si un entier p-adique $x =$ (x,,) n'appartient pas à $p\mathbf{Z}_p$, chacune de ses composantes x_m est inversible dans le facteur correspondant $\mathbf{Z}/p^m\mathbf{Z}$ et (x_m^{-1}) est inverse de x dans $\mathbf{Z}_p$; ainsi l'idéal maximal $p\mathbf{Z}_p$ est exactement l'ensemble des éléments non inversibles de $\mathbf{Z}_p$, et c'est donc le seul idéal maximal : l'anneau $\mathbf{Z}_p$ est *local* (cf. ANNEAUX ET ALGÈBRES) etsoncorpsrésiduel est $\mathbf{Z}_p/p\mathbf{Z}_p \simeq \mathbf{F}_p$, corps à p éléments. Les puissances successives de l'idéal maximal forment une suite décroissante $(p^m\mathbf{Z}_p)$ d'idéaux dont l'intersection est visiblement {0) ; le plus grand de ces idéaux est $p^0\mathbf{Z}_p = \mathbf{Z}_p$. La multiplication par p^n est injective dans $\mathbf{Z}_p$; il suffit de le vérifier pour $n = 1$, et $px = 0$ équivaut à $px_m = 0$ dans $\mathbf{Z}/p^m\mathbf{Z}$ pour tout m, ce qui donne :

$$x_m \in p^{m-1}\mathbf{Z}/p^m\mathbf{Z},$$

d'où x_m , = 0. L'anneau $\mathbf{Z}_p$ s'applique donc bijectivement sur l'idéal $p^n\mathbf{Z}_p$, et l'ensemble $U = \mathbf{Z}_p - p\mathbf{Z}_p$ de ses éléments inversibles s'applique bijectivement sur :

$$p^n\mathbf{Z}_p - p^{n+1}\mathbf{Z}_p = p^n\mathrm{U}\,;$$

on voit ainsi que l'ensemble Z_p {0} est réunion disjointe des p^nU ($n \in$ N). Autrement dit, tout entier p-adique non nul x s'écrit ***d'une seule manière*** sous la forme $p^n u$ avec $n \in$ N et $u \in$ U entier p-adique inversible ; l'entier naturel n s'appelle la ***valuation p-adique*** de x et se note $v_p(x)$. Si x $= p^m u$ et y $= p^n v$ sont des entiers p-adiques non nuls, avec $u \in$ U et u $\in$ U, on a :

$$xy = p^{m+n}uv \neq 0,$$

car uv est encore inversible, donc $\mathbf{Z}_p$ est un ***anneau intègre*** ; on voit en même temps que :

$$v_p(xy) = v_p(x) + v_p(y),$$

et on peut vérifier par ailleurs l'inégalité :

$$v_p(x+y) \geqslant \inf(v_p(x), v_p(y));$$

ces deux propriétés de la valuation p-adique restent vraies même si x ou y est nul lorsque l'on pose $v_p(0) = +\infty$, élément abstrait ajouté à N avec les propriétés habituelles : on a $n < +\infty$ et $n + (+\infty) = +\infty$ pour tout $n \in$ N.

Considérons un idéal non nul a de $\mathbf{Z}_p$; si n est la borne inférieure des valuations des éléments de a, on voit sans peine que a est l'idéal engendré par p^n. Ainsi les seuls idéaux de $\mathbf{Z}_p$ sont (0) et les $p^n\mathbf{Z}_p$; en particulier $\mathbf{Z}_p$ est un ***anneau principal*** et il n'a qu'un ***seul idéal premier non nul*** $p\mathbf{Z}_p$: un tel anneau s'appelle un ***anneau de valuation discrète*** (cf. ANNEAUX COMMUTATIFS). La décomposition en facteurs premiers d'un entierp-adique x est de la forme $x = p^n u$ avec $n = v_p(x)$ et $u \in$ U.

Il y a une autre structure intéressante sur l'anneau $\mathbf{Z}_p$. On peut en effet considérer chaque $\mathbf{Z}/p^m\mathbf{Z}$ comme un anneau topologique discret et munir :

$$\prod \mathbf{Z}/p^m\mathbf{Z}$$

de la topologie produit ; on obtient ainsi un anneau topologique compact, comme produit d'ensembles compacts (cf. théorème de Tychonoff, in TOPOLOGIE GÉNÉRALE). Comme les applications canoniques φ_{km} sont continues, $\mathbf{Z}_p$ est une partie fermée de cet anneau produit ; c'est donc encore un ***anneau topologique compact*** pour la topologie induite. Un système fondamental de voisinages de 0 pour la topologie produit est formé par les ensembles :

$$V_n = \{(x_m) \mid x_m = 0,\ m < n\};$$

la trace de V_n sur $\mathbf{Z}_p$ n'est autre que l'idéal $p^n\mathbf{Z}_p$ qui est encore égal à l'ensemble des entiers p-adiques de valuation $\geqslant n$. Pour tout entier p-adique x, on pose :

$$|x|_p = p^{-v_p(x)};$$

ce dernier nombre s'appelle ***valeur absolue p-adique*** de x, et l'on a :

$$|xy|_p = |x|_p\,|y|_p,\quad |x+y|_p \leqslant \sup(|x|_p, |y|_p),$$

avec $|x|_p = 0$ si et seulement si $x = 0$. On voit alors que :

$$(x, y) \mapsto d(x, y) = |y - x|_p$$

est une ***distance*** (ultramétrique) sur $\mathbf{Z}_p$ (cf. espaces MÉTRIQUES) ; les remarques précédentes montrent que cette distance définit la topologie de $\mathbf{Z}_p$. Notons maintenant que les isomorphismes :

$$\mathbf{Z}_p/p^n\mathbf{Z}_p \simeq \mathbf{Z}/p^n\mathbf{Z}$$

montrent que ***Z*** est partout dense dans $\mathbf{Z}_p$. Comme $\mathbf{Z}_p$ est complet (puisque compact), il est isomorphe au ***complété*** de ***Z*** pour la distance p-adique $|y - x|_p$.

Soit ***S*** un ***système de représentants*** de $\mathbf{F}_p$ dans $\mathbf{Z}_p$, c'est-à-dire un ensemble d'entiers padiques possédant exactement un élément dans chaque classe mod $p\mathbf{Z}_p$. Par exemple, on peut prendre :

$$S = \{0, 1, \ldots, p-1\};$$

mais nous verrons plus loin un autre système de représentants plus intéressant. Pour tout entier p-adique x, il existe un élément s_0 de S et un seul congru à x mod $p\mathbf{Z}_p$; on a $x - s_0 = px_1$ avec $x_1 \in \mathbf{Z}_p$. De même, il existe un couple unique d'éléments $s_1 \in$ S et $x_2 \in \mathbf{Z}_p$ tels que $x_1 = s_1 + px_2$, *ce* qui donne $x = s_0 + ps_1 + p^2x_2$; en raisonnant par récurrence, on trouve pour tout n une manière unique d'écrire x sous la forme :

$$x = s_0 + s_1 p + ... + s_n p^n + x_{n+1} p^{n+1},$$

avec $s_i \in$ S et $x_{n+1} \in \mathbf{Z}_p$. Lorsque n tend vers l'infini, le reste $x_{n+1}p^{n+1}$ tend vers 0 dans $\mathbf{Z}_p$; on trouve donc un développement en série infinie :

$$x = \sum_i s_i p^i,$$

avec des coefficients $s_i \in$ S ; inversement, toute série de cette forme converge dans $\mathbf{Z}_p$ et définit un entier p-adique, comme on le voit, en appliquant le critère de Cauchy ; notons d'ailleurs que, d'après l'inégalité ultramétrique :

$$d(x,z) \leqslant \sup(d(x,y), d(y,z)),$$

une série converge dans $\mathbf{Z}_p$ dès que son terme général tend vers 0.

Au début du siècle, K. Hensel a introduit les nombres p-adiques en les définissant par des développements en série du type précédent.

Le corps des fractions $\mathbf{Q}_p$ de $\mathbf{Z}_p$ s'appelle le *corps des nombres p-adiques* ; il contient le corps Q des nombres rationnels comme sous-corps (autrement dit, c'est un corps de caractéristique 0). Chaque nombre p-adique peut s'écrire comme une fraction x/p^n avec au numérateur un entier p-adique x et au dénominateur une puissance de p ; la valuation p-adique se prolonge à $\mathbf{Q}_p$ en posant :

$$v_p\left(\frac{x}{p^n}\right) = v_p(x) - n \in \mathbf{Z} \cup \{+\infty\}.$$

On a encore :

$$v_p(xy) = v_p(x) + v_p(y),$$
$$v_p(x+y) \geqslant \inf(v_p(x), v_p(y)),$$

pour $x, y \in$ Q,. On fait de $\mathbf{Q}_p$ un *corps topologique localement compact* en le munissant de la topologie pour laquelle $\mathbf{Z}_p$ est un sous-groupe additif ouvert (cf. algèbre TOPOLOGIQUE) ; cette topologie peut aussi être définie par la distance ultramétrique $(x, y) \mapsto d(x, y) = |y - x|_p$ correspondant à la valeur absolue $|x|_p = p^{-v_p(x)}$. Remarquons que $\mathbf{Z}_p$ n'est autre que la boule unité :

$$\mathbf{Z}_p = \{x \in \mathbf{Q}_p \mid |x|_p \leqslant 1\} = \{x \mid v_p(x) \geqslant 0\};$$

de même, l'idéal maximal $p\mathbf{Z}_p$ est défini par l'inégalité $|x|_p < 1$ ou aussi bien $v_p(x) > 0$. Le corps $\mathbf{Q}_p$ est *complet,* car il est localement compact, et Q est un sous-corps partout dense ; donc $\mathbf{Q}_p$ est isomorphe au complété de Q pour la distance p-adique.

2. Équations p-adiques ; lemme de Hensel

Revenons aux considérations du début et étudions un système d'équations :

$$f_\alpha(x_1, x_2, ..., x_m) = 0,$$

où $\alpha = 1, 2, ..., \nu$ et où les f_α sont des polynômes à coefficients dans $\mathbf{Z}_p$; on cherche les solutions $(x_1, x_2, ..., x_m)$ dans $(\mathbf{Z}_p)^m$. Par réduction modulo p^n, on en déduit un système d'équations $f_{\alpha,n} = 0$ dans $\mathbf{Z}/p^n\mathbf{Z}$. Pour que le système étudié ait une solution dans $(\mathbf{Z}_p)^m$, il faut et il suffit que pour tout n le système réduit mod p^n

ait une solution dans $(\mathbf{Z}/p^n\mathbf{Z})^m$. En effet, une solution dans $(\mathbf{Z}_p)^m$ n'est autre qu'une suite de solutions mod p^n pour tous les n, qui se correspondent par les applications canoniques φ_{kn} ; si X,, C $(\mathbf{Z}/p^n\mathbf{Z})^m$ désigne l'ensemble des solutions mod p^n, on voit que l'image dans X_n de l'image X des solutions dans $(\mathbf{Z}_p)^m$ est :

$$Y_n = \bigcap_{q \geqslant n} \varphi_{nq}(X_q),$$

qui est non vide si chaque X, est non vide (intersection décroissante d'ensembles finis non vides ; en termes plus savants, on observe que X est limite projective des X_n et que la limite projective d'une suite d'ensembles finis non vides est non vide).

De la même manière, on prouve que l'existence d'une solution *primitive* dans $(\mathbf{Z}_p)^m$ équivaut à l'existence d'une solution primitive mod p^n pour tout n ; un élément primitif de $(\mathbf{Z}_p)^m$, ou de $(\mathbf{Z}/p^n\mathbf{Z})^m$, est, par définition, un élément dont l'une des m coordonnées est inversible. Remarquons que l'existence d'une solution primitive dans $(\mathbf{Z}_p)^m$ équivaut encore à l'existence d'une solution différente de 0 dans $(\mathbf{Q}_p)^m$; on le voit en réduisant au même dénominateur.

Le résultat suivant donne une condition suffisante pour qu'un zéro mod p^n d'un polynôme $f \in \mathbf{Z}_p[X_1, X_2, \ldots, X_m]$ se relève en un zéro dans $\mathbf{Z}_p$. Soit $x \in (\mathbf{Z}_p)^m$ tel que f(x) $\equiv$ 0 (mod p^n) ; s'il existe un indice j tel que :

$$\left|\frac{\partial f}{\partial X_j}(x)\right|_p > p^{-n/2},$$

il existe un élément $y \in (\mathbf{Z}_p)^m$ tel que $f(y) = 0$ et $y \equiv x$ (mod p^{n-k}), où :

$$k = v_p\left(\frac{\partial f}{\partial X_j}(x)\right).$$

On construit y comme limite d'une suite (x,) d'éléments de $(\mathbf{Z}_p)^m$ vérifiant :

$$(x_q) \equiv 0 (\text{mod}\, p^{n+q}),\ v_p\left(\frac{\partial f}{\partial X_j}(x_q)\right) = k,$$
$$x_0 = x,\ x_{q+1} \equiv x_q (\text{mod}\, p^{n+q-k});$$

l'existence d'une telle suite résulte du lemme suivant.

Lemme. Soit $f \in \mathbf{Z}_p[X]$ et $x \in \mathbf{Z}_p$ tels que f(x) $\equiv$ 0 (mod p^n). Si :

$$|f'(x)|_p > p^{-n/2},$$

il existe $x' \in$ Z,, tel que :

$$f(x') \equiv 0\ (\text{mod}\, p^{n+1}),$$
$$|f'(x')|_p = |f'(x)|_p,\ x' \equiv x\ (\text{mod}\, p^{n-k}),$$

où k = $v_p(f'(x))$.

Pour démontrer ce lemme, on cherche x' sous la forme $x' = x + p^{n-k}z$ et on utilise la formule de Taylor :

$$f(x + p^{n-k}z) = f(x) + p^{n-k}zf'(x) + p^{2n-2k}t,$$

où test un entierp-adique ; disons que $f(x)$ est multiple de p^n et que la valuation de f'(x) est k, soit f'(x) = $p^n y$ et f"(x) $= p^k u$ avec $y \in \mathbf{Z}_p$ et $u \in$ U, donc :

$$f(x + p^{n-k}z) = p^n(y + zu + p^{n-2k}t).$$

Comme par hypothèse $n - 2k \geqslant 1$, il suffit de choisir z tel que y $+\ zu$ soit multiple de p, c'est-à-dire $z \equiv -yu^{-1}$ (modp) ; alors $f(x + p^{n-k}z)$ est divisible par p^{n+1} et :

$$'(x + p^{n-k}z) = f'(x) + p^{n-k}zt'$$
$$= p^k(u + p^{n-2k}zt')$$

a pour valuation k.

Appliquons ce résultat dans le cas où n = 1 ; on a alors nécessairement $k = 0$ et on voit que, si $\xi \in (\mathbf{F}_p)^m$ est un zéro du polynôme :

$$\bar{f} \in \mathbf{F}_p[X_1, X_2, \ldots, X_m],$$

déduit de f par réduction modulo p, et si une au moins des dérivées partielles premières de $\bar{f}$ ne s'annule pas en ξ (zéro « simple »), alors il existe un zéro de f dans $(\mathbf{Z}_p)^m$ qui relève ξ. Par exemple, si f est homogène de degré 2 (forme quadratique) et si son discriminant est inversible dans $\mathbf{Z}_p$, $\bar{f}$ est une forme quadratique non dégénérée à coefficients dans $\mathbf{F}_p$; lorsque $p \neq 2$ tout $\xi \in (\mathbf{F}_p)^m - \{0\}$ tel que $\bar{f}(\xi) = \bar{a}$ est une racine simple et se relève donc en un $x \in (\mathbf{Z}_p)^m$ tel que $f(x) = a$. Lorsque $p = 2$, on montre de même que, si x est un élément primitif de $(\mathbf{Z}_2)^m$ tel que $f(x) \equiv a \pmod 8$, il existe $y \in (\mathbf{Z}_2)^m$ tel que $f(y) = a$ et $y \equiv x \pmod 4$. Pour $m = 1$, ces résultats permettent de déterminer les éléments inversibles de $\mathbf{Z}_p$ qui sont les carrés ; lorsque $p \neq 2$, ce sont les éléments de U dont la classe mod p est un carré dans $\mathbf{F}_p$; nous étudierons plus loin les carrés de $\mathbf{Z}_p$ par une autre méthode.

Le lemme de Hensel est un résultat voisin du précédent et s'énonce ainsi : soit $f \in \mathbf{Z}_p[X]$ un polynôme à une indéterminée à coefficients entiers p-adiques ; supposons donnée un décomposition $\bar{f} = \varphi\psi$ de la réduction de f modulo p en produit d'un polynôme unitaire φ de degré d et d'un polynôme ψ étranger à φ ; il existe un couple unique (g, h) de polynômes g, h appartenant à $\mathbf{Z}_p[X]$ tels que $f = gh$, $\bar{g} = \varphi$, $\bar{h} = \psi$ et que g soit unitaire de degré d.

3. Structure du groupe multiplicatif $\mathbf{Q}_p^*$

On sait que tout élément non nul de $\mathbf{Q}_p$ s'écrit d'une seule manière sous la forme $p^n u$ avec $n \in \mathbf{Z}$ et $u \in$ U, groupe des éléments inversibles de $\mathbf{Z}_p$; cela donne immédiatement un isomorphisme $\mathbf{Q}_p^* \simeq \mathbf{Z} \times$ U. Il reste à étudier la structure du groupe U ; on définit une *filtration décroissante* (U_n) de U en posant pour tout $n > 0$:

$$U_n = \{x \in \mathbf{Z}_p \mid x \equiv 1 \pmod{p^n}\}, \quad U_0 = U;$$

ainsi U_n est le noyau de l'homomorphisme canonique de U dans le groupe des éléments inversibles de $\mathbf{Z}/p^n\mathbf{Z}$. Il est clair que $U/U_1 \simeq \mathbf{F}_p^*$, groupe cyclique d'ordre $p - 1$; pour $n \geqslant 1$, une bijection $x \mapsto 1 + p^n x$ de $\mathbf{Z}_p$ sur U_n applique $p\mathbf{Z}_p$ sur U_{n+1} et définit un isomorphisme de groupes :

$$\mathbf{Z}/p\mathbf{Z} \sim \mathbf{Z}_p/p\mathbf{Z}_p \rightrightarrows U_n/U_{n+1},$$

en vertu de l'identité :

$$(1 + p^n x)(1 + p^n y) = 1 + p^n(x + y) + p^{2n}xy \equiv 1 + p^n(x + y) \pmod{p^{n+1}};$$

ainsi U_n/U_{n+1} est cyclique d'ordre p pour $n \geqslant 1$.

De ces considérations on peut déduire qu'il existe un sous-groupe unique V de U isomorphe à $\mathbf{F}_p^*$ et que U est isomorphe au produit V X U_1 ; le sous-groupe V est l'ensemble des entiers p-adiques x tels que $x^{p-1} = 1$, et V $\cup \{0\}$ est un système de représentants de $\mathbf{F}_p$ dans $\mathbf{Z}_p$ qui est stable par multiplication, c'est un système de représentants multiplicatifs (cf. chap. 1). Pour obtenir ces résultats, on considère U et U_1 comme limites projectives des suites de groupes (U/U_n) et (U_1/U_n) respectivement, et on est ramené à prouver l'existence d'un unique sous-groupe V_n de U/U_n isomorphe à $\mathbf{F}_p$ et tel que :

$$U/U_n \simeq V_n \times U_1/U_n;$$

comme U_1/U_n est d'ordre p^{n-1} et que :

$$(U/U_n)/(U_1/U_n) \simeq \mathbf{F}_p^*$$

est d'ordre $p - 1$ premier à p^{n-1}, on peut démontrer que U/U_n est produit de U_1/U_n par le sous-groupe des racines

$(p-1)$-ièmes de 1 en utilisant l'identité de Bezout (cf. ANNEAUX COMMUTATIFS). Chemin faisant, nous avons démontré que le corps des nombres p-adiques $\mathbf{Q}_p$ contient les racines $(p-1)$-ièmes de 1.

Il faut enfin élucider la structure du groupe U_1. Nous allons voir que U_1 est isomorphe au groupe additif $\mathbf{Z}_p$ si p est différent de 2, et à $\{\pm 1\} \times \mathbf{Z}_2$ si $p = 2$. Pour $p \neq 2$, on choisit un élément a de U_1 qui n'appartient pas à U_2 et on considère l'homomorphisme $x \mapsto a^x$ de Z dans U_1 ; on a $a = 1 + pu$ avec $u \in U$, donc $a^x = 1 + xpu + p^2 t$ (t entier p-adique) par le développement du binôme, et, pour x premier à p, cela montre que a^x est encore un élément de U_1 qui n'est pas dans U_2 ; au contraire, pour $x = p^h$, on trouve que a^x est un élément de U_{h+1} qui n'est pas dans U_{h+2}, en remarquant que pour tout n la puissance p-ième d'un élément de $U_n - U_{n+1}$ appartient à $U_{n+1} - U_{n+2}$. Ces résultats permettent de voir que l'image réciproque de U_{n+1} dans Z est exactement $p^n\mathbf{Z}$; par conséquent, $x \mapsto a^x$ définit un homomorphisme injectif de $\mathbf{Z}/p^n\mathbf{Z}$ dans U_1/U_{n+1}, et cet homomorphisme est même un isomorphisme, car les deux groupes ont le même nombre d'éléments. En passant à la limite projective pour $n \to \infty$, on obtient l'isomorphisme cherché $\mathbf{Z}_p \rightrightarrows U_1$; on notera encore a^x l'image d'un entier p-adique x par cet isomorphisme. Dans le cas où $p = 2$, on observe d'abord que :

$$U = U_1 \simeq \{\pm 1\} \times U_2,$$

et on définit un isomorphisme de $\mathbf{Z}_2$ sur U_2 à partir de l'homomorphisme $x \mapsto a^x$ de Z dans U_2 construit à l'aide d'un élément a de U_2 qui n'appartient pas à U_3 (ainsi a est congru à 5 mod 8). En résumé, on a trouvé que le groupe multiplicatif $\mathbf{Q}_p^*$ est isomorphe à :

$$\mathbf{Z} \times (\mathbf{Z}/(p-1)\mathbf{Z}) \times \mathbf{Z}_p,$$

si $p \neq 2$ et à :

$$\mathbf{Z} \times (\mathbf{Z}/2\mathbf{Z}) \times \mathbf{Z}_2,$$

si $p = 2$.

Nous sommes maintenant en mesure de déterminer quels sont les carrés dans $\mathbf{Q}_p$. Pour $p \neq 2$, on a :

$$\mathbf{Q}_p^{*2} \simeq 2\mathbf{Z} \times \mathbf{F}_p^{*2} \times \mathbf{Z}_p,$$

car, 2 étant inversible dans $\mathbf{Z}_p$, on a $2\mathbf{Z}_p = \mathbf{Z}_p$; on retrouve le fait qu'un élément inversible de $\mathbf{Z}_p$ est un carré si et seulement si son image dans $\mathbf{F}_p$ est un carré. Le groupe quotient $\mathbf{Q}_p/\mathbf{Q}_p^{*2}$ est isomorphe à :

$$(\mathbf{Z}/2\mathbf{Z}) \times (\mathbf{F}_p/\mathbf{F}_p^{*2}) \simeq (\mathbf{Z}/2\mathbf{Z}) \times (\mathbf{Z}/2\mathbf{Z}) ;$$

c'est un groupe à 4 éléments qui admet pour système de représentants dans $\mathbf{Q}_p^*$ l'ensemble $\{1, p, u, up\}$ où u est un entier qui n'est pas résidu quadratique mod p.

Lorsque $p = 2$, on a :

$$\mathbf{Q}_2^{*2} \simeq 2\mathbf{Z} \times 0 \times 2\mathbf{Z}_2 \simeq 2\mathbf{Z} \times U_3,$$

et alors :

$$U_3 = \{u \in \mathbf{Z}_2 \quad u \equiv 1 \pmod 8\}$$

est l'ensemble des carrés inversibles de $\mathbf{Z}_2$;

$$\mathbf{Q}_2/\mathbf{Q}_2^{*2} \simeq \mathbf{Z}/2\mathbf{Z} \to \mathbf{Z}/2\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$$

est un groupe d'ordre 8 qui admet pour système de représentants $\{1, 2, 3, 5, 6, 7, 10, 14\}$ dans $\mathbf{Q}_2^*$. Notons que $\mathbf{Q}_p^{*2}$ contient U_1 si $p \neq 2$ et U_3 si $p = 2$; donc, c'est un sous-groupe *ouvert* de $\mathbf{Q}_p^*$; le groupe quotient $\mathbf{Q}_p^*/\mathbf{Q}_p^{*2}$ peut être considéré comme un espace vectoriel sur le corps à 2 éléments $\mathbf{F}_2$, de dimension 2 ou 3 suivant que *$p \neq 2$ ou $p = 2$*. Sur cet espace vectoriel, il y a une forme bilinéaire canonique, qui est symétrique et non dégénérée ; elle est définie par le *symbole de Hilbert* (a, b), $a, b \in \mathbf{Q}_p^*$, qui vaut 1 ou -1

suivant qu'il existe ou non un élément non nul de $(\mathbf{Q}_p)^3$ qui annule la forme quadratique $Z^2 \quad aX^2 \; bY^2$ (cf. DIVISIBILITÉ ; la partie C ci-après -Nombres algébriques ; formes QUADRATIQUES). À l'aide du symbole de Hilbert, on définit un invariant $\varepsilon(f)$ associé à toute forme quadratique f à coefficients dans $\mathbf{Q}_p$; si :

$$f(x_1, x_2, \ldots, x_m) = a_1 x_1^2 + a_2 x_2^2 + \quad + a_m x_m^2,$$

dans une base orthogonale, on a :

$$\varepsilon(f) = \prod_{i<j} @,,a,)$$

et on peut montrer que $\varepsilon(f)$ ne dépend pas de la base orthogonale choisie. Pour que deux formes quadratiques à coefficients dans $\mathbf{Q}_p$ soient isomorphes, il faut et il suffit qu'elles aient même rang, même invariant ε et que leurs discriminants aient la même classe dans $\mathbf{Q}_p^*/\mathbf{Q}_p^{*2}$; on trouve ainsi 4 (resp. 8) classes de formes de rang 1 et 7 (resp. 15) classes de formes de rang 2 et enfin 8 (resp. 16) classes dc formes de rang $n \geqslant 3$ si $p \neq 2$ (resp. $p = 2$).

4. Analyse p-adique

On peut développer une théorie des fonctions analytiques de variables p-adiques en définissant de telles fonctions par des développements en séries entières convergentes (cf. FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe).

Par exemple, la série *exponentielle* :

$$\exp X = 1 + x + \quad + \frac{X^n}{n!} +$$

converge dans le « disque ouvert » de $\mathbf{Q}_p$ défini par l'inégalité :

$$v_p(x) > \frac{1}{p-1};$$

en effet :

$$v_p(n!) = v_p(1) + v_p(2) + \quad + v_p(n),$$

et le nombre d'entiers $k \leqslant n$ tels que $v_p(k) \geqslant r$ est égal à la partie entière $[n/p^r]$ du nombre rationnel n/p^r, ce qui donne :

$$v_p(n!) = \frac{n}{p} + \frac{n}{p^2} + \ldots = \frac{n - s_n}{p-1},$$

en désignant par s_n la somme des coefficients du développement p-adique de n (si $n = \Sigma a_i p^i$ avec $0 \leqslant a_i \leqslant p-1$, on a $s_n = \Sigma a_i$) ; on en déduit facilement que $v_p(1/n!)/n$ converge vers $1/(p \quad 1)$ pour n infini, c'est-à-dire que $|1/n!|^{1/n}$ converge vers $p^{1/(p \quad 1)}$. On peut donc définir une fonction exponentielle $x \mapsto \exp x$ pour $|x| < p^{-1/(p-1)}$; il est clair que l'on a :

$$\exp(x + y) = \exp x \,.\, \exp y$$

et que la fonction exponentielle ne s'annule pas dans son disque de convergence : elle définit un homomorphisme de ce disque, qui est un sous-groupe du groupe additif de $\mathbf{Q}_p$ (à savoir $p\mathbf{Z}_p$ si $p \neq 2$ et $4\mathbf{Z}_2$ si $p = 2$), dans le groupe multiplicatif de $\mathbf{Q}_p^*$. C'est en fait un isomorphisme sur un sous-groupe de $\mathbf{Q}_p^*$, comme on peut le montrer à l'aide de la fonction logarithme définie par la série :

$$\log(1 + Y) = Y - \frac{Y^2}{2} + \quad + (-1)^{n-1} \frac{Y^n}{n} +$$

qui converge pour $v_p(y) > 0$ et définit donc un logarithme dans U, ; on peut vérifier que :

$$\log(\exp x) = x, \quad \exp(\log(1 + x)) = 1 + x$$

pour $v_p(x) > 1/(p-1)$ et, par suite, exp définit un isomorphisme de $p\mathbf{Z}_p$ sur U, si $p \neq 2$ et de $4\mathbf{Z}_2$ sur U_2 si $p = 2$: on retrouve les isomorphismes du chapitre 3.

Comme le corps $\mathbf{Q}_p$ est totalement discontinu, on ne peut espérer une théorie globale raisonnable pour les fonctions analytiques au sens habituel, c'est-à-dire

définies localement par des développements en série. Il y a cependant une théorie globale pour les fonctions « strictement holomorphes » ; donnons, par exemple, la définition des fonctions strictement holomorphes dans la « couronne » :

$$\{x \in \mathbf{Q}_p \mid r \leqslant |x| \leqslant \mathrm{R}\}.$$

Ce sont les fonctions définies par des développements de Laurent :

$$f(x) = \sum_{n \in \mathbf{Z}} a_n x^n,$$

qui vérifient la condition de convergence suivante : pour tout $\rho \in [r, \mathrm{R}]$, $a_n \rho^n$ tend vers 0 lorsque n tend vers $\pm\infty$; l'espace $\mathrm{L}(r, \mathrm{R})$ de ces développements de Laurent est un espace de Banach sur le corps $\mathbf{Q}_p$ pour la norme (ultramétrique) :

$$\|f\| = \sup(\sup_n |a_n| r^n, \sup_n |a_n| \mathrm{R}^n),$$

et on peut y définir une multiplication qui en fait une algèbre de Banach. On démontre que $\mathrm{L}(r, \mathrm{R})$ est un anneau principal et que ses éléments irréductibles sont les polynômes unitaires irréductibles dont les racines (dans une clôture algébrique de $\mathbf{Q}_p$) ont des valeurs absolues dans l'intervalle $[r, \mathrm{R}]$, ainsi que les produits de ces polynômes par des éléments inversibles ; ces derniers sont les développements de la forme :

$$cx^{\mathrm{N}}\Big(1 + \sum_{n \geqslant 1} a_n x^n\Big),$$

avec $a_n r^n < 1$ et $a_n \mathrm{R}^n < 1$ pour tout n. M. Lazard a généralisé ces résultats au cas d'une couronne « ouverte » (définie par des inégalités strictes) et a obtenu des théorèmes analogues à ceux de Weierstrass (développement en produit infini d'une fonction strictement méromorphe) et de Mittag-Leffler.

En s'inspirant de la théorie de Jacobi, J. Tate a élaboré une théorie analytique des fonctions elliptiques sur un corps p-adique. Soit $q \in \mathbf{Q}_p$ tel que $0 < |q| < 1$, on considère le corps des fonctions strictement méromorphes dans $\mathbf{Q}_p - \{0\}$ (couronne de rayons 0 et ∞) qui sont invariantes par la multiplication par q, c'est-à-dire $f(qx) = f(x)$; on peut montrer que c'est un corps de fonctions algébriques d'une variable et que son genre est 1, c'est-à-dire qu'il s'identifie au corps des fonctions rationnelles sur une courbe elliptique. Cette théorie donne l'uniformisation de certaines courbes elliptiques sur $\mathbf{Q}_p$; plus récemment, M. Raynaud et D. Mumford ont étudié d'une manière analogue les variétés abéliennes et les courbes algébriques de genre $\geqslant 2$.

J. Tate a également donné une définition des espaces analytiques p-adiques qui permet d'étudier les fonctions analytiques de plusieurs variables. Les principaux théorèmes de la théorie des faisceaux cohérents ont été établis pour ces espaces (J. Tate et R. Kiehl). Le renouveau d'intérêt pour l'analyse p-adique vient surtout de la démonstration par Dwork de la rationalité de la fonction zêta d'une variété algébrique sur un corps fini (cf. fonction ZÊTA) ; la méthode utilisée consiste à démontrer la rationalité d'une série formelle à coefficients entiers en étudiant ses propriétés comme fonction analytique p-adique (pour différents nombres premiers p) ; les propriétés en question s'obtiennent, dans le cas de la fonction zêta, en appliquant la théorie de Fredholm p-adique (J.-P. Serre ; cf. théorie SPECTRALE).

Il existe aussi une théorie des groupes de Lie p-adiques tout à fait analogue à celle des groupes de Lie réels ou complexes (cf. GROUPES - Groupes de Lie). Le résultat du chapitre 3 sur la structure de $\mathbf{Q}_p^*$ se généralise ainsi : tout groupe de Lie

commutatif de dimension n sur $\mathbf{Q}_p$ contient un sous-groupe ouvert isomorphe à $(\mathbf{Z}_p)^n$; si le groupe considéré est compact, c'est une extension d'un groupe fini par $(\mathbf{Z}_p)^n$. La théorie des représentations linéaires pour les groupes de Lie p-adiques a été développée par F. Bruhat ; dans le cas commutatif, on dispose de la transformation de Fourier comme dans le cas classique et le groupe dual du groupe additif $\mathbf{Q}_p$ est isomorphe à $\mathbf{Q}_p$. Signalons enfin l'existence d'une théorie des fonctions sphériques p-adiques et les travaux de F. Bruhat et J. Tits sur la structure des groupes algébriques p-adiques.

5. Extensions

On connaît beaucoup d'autres anneaux de valuation discrète que les anneaux $\mathbf{Z}_p$; nous pouvons citer l'anneau $k[[T]]$ des séries formelles à une indéterminée à coefficients dans un corps k, ou l'anneau local d'un point régulier sur une courbe algébrique (ou sur une courbe analytique complexe ; cf. GÉOMÉTRIE ALGÉBRIQUE). Si A est un anneau de valuation discrète de corps des fractions K et si πA est l'unique idéal premier non nul de A, les idéaux de A sont 0 et les π^nA ($n \in$ N) ; tout élément $x \neq 0$ de K s'écrit d'une seule manière sous la forme $x = \pi^n u$ où $n \in$ Z et où u est un élément inversible de A ; la *valuation* de x est l'entier $v(x) = n$ et l'application v : K* $\rightarrow$ Z est un homomorphisme surjectif de groupes vérifiant l'inégalité $v(x + y) \geqslant \inf\ (v(x), v(y))$. Inversement, la donnée d'une valuation discrète u : K* $\rightarrow$ Z détermine un sous-anneau :

$$\mathrm{A} = \{x \in \mathrm{K} \mid v(x) \geqslant 0\},$$

où l'on a posé u(0) = + ∞, qui est un anneau de valuation discrète et dont l'idéal maximal est l'ensemble des éléments de valuation > 0. La valuation définit une valeur absolue ultramétrique :

$$x \mapsto |x| = a^{v(x)},$$

où a est un nombre réel fixé appartenant à l'intervalle]0, 1] ; on a :

$$|xy| = |x| . |y|, \quad |x + y| \leqslant \sup(|x|, |y|),$$

et x ne s'annule que pour x = 0. Dans le cas où K est complet pour la topologie définie par cette valeur absolue, il possède des propriétés très semblables à celles de $\mathbf{Q}_p$; pour qu'il soit localement compact, il faut et il suffit qu'il soit complet et que le corps résiduel k = A/πA soit fini (cf. la partie C ci-après - Nombres algébriques). Si A est un anneau de valuation discrète complet et si L est une extension finie de son corps des fractions K, on démontre que la fermeture intégrale B de A dans L est encore un anneau de valuation discrète complet et que c'est un A-module libre de rang [L :K] ; désignons par π un générateur de l'idéal maximal de A (une «uniformisante ») et par w la valuation définie par B ; l'entier $e = w(\pi)$ s'appelle l'indice de ramification de L sur K. La valeur absolue de K se prolonge d'une manière unique à L. Considérons un anneau de valuation discrète complet A ; supposons que son corps des fractions K soit de caractéristique 0 et son corps résiduel k de caractéristique p > 0. Alors. l'injection canonique de Z dans A (resp. de Q dans K) se prolonge par continuité en :

$$\mathbf{Z}_p \rightarrow \mathrm{A} \quad (\text{resp. } \mathbf{Q}_p \rightarrow \mathrm{K})\ ;$$

l'entier $e = v(p)$, où v est la valuation définie par A, s'appelle l'indice de ramification absolue de A. On démontre (I. S. Cohen) que, pour tout corps *parfait* k de caractéristique p, il existe un anneau de

valuation discrète complet A absolument non ramifié (c'est-à-dire d'indice de ramification absolue égal à 1, ce qui signifie que l'idéal maximal de A est p A) dont le corps résiduel est k; cet anneau est unique à isomorphisme (unique) près, et sa construction se fait au moyen des *vecteurs de Witt;* ainsi $\mathbf{Z}_p$ est l'anneau de valuation discrète complet absolument non ramifié de corps résiduel $\mathbf{F}_p$.

Si A est un anneau de valuation discrète complet de caractéristique un nombre premier p, alors son corps résiduel k est aussi de caractéristique p, et on démontre qu'il admet dans A un système de représentants qui est un sous-corps (les représentants multiplicatifs sont également additifs ; cf. chap. 3) ; on en déduit, en utilisant des développements de type hensélien par rapport aux puissances d'une uniformisante, que A est isomorphe à l'anneau des séries formelles k[[T]]. Il en est de même si k est de caractéristique 0.

L'analyse p-adique se généralise en remplaçant $\mathbf{Q}_p$ par un corps valué complet ultramétrique quelconque. Les résultats et les méthodes sont les mêmes (sauf ceux qui font intervenir les propriétés arithmétiques particulières à $\mathbf{Q}_p$).

CHRISTIAN HOUZEL

Bibliographie

Voir la bibliographie à la fin de la partie C ci-après.

C. Nombres algébriques

Les mathématiciens grecs avaient découvert que certains rapports de grandeurs ne sont pas rationnels, c'est-à-dire qu'ils ne sont pas égaux au rapport de deux entiers : il en est ainsi du rapport de la diagonale d'un carré à son côté, puisque aucun nombre rationnel n'a un carré égal à 2. Plus généralement, Théétète (Ve s. avant J.-C.) a établi qu'un entier qui n'est pas le carré d'un entier n'est pas non plus le carré d'un nombre rationnel. Le dixième livre des ***Éléments*** d'Euclide est consacré à l'étude et à la classification des grandeurs irrationnelles rencontrées dans les constructions géométriques.

Les recherches sur les équations algébriques ont toujours été inséparables de problèmes touchant la nature des solutions de ces équations. Durant le XVIIIe siècle, il fut établi que les n racines d'une équation algébrique de degré n à coefficients réels étaient des nombres complexes (cf. nombres COMPLEXES). On appelle maintenant *nombre algébrique* tout nombre complexe qui est racine d'une équation algébrique à coefficients *rationnels* : ainsi $\sqrt{2}$, racine de l'équation $x^2 - 2 = 0$, ou bien i, racine de l'équation $x^2 + 1 = 0$, ou encore $e^{2i\pi/n}$, racine de $x^n - 1 = 0$, sont des nombres algébriques ; au contraire e, π, log 2 ou i^i ne sont pas des nombres algébriques (cf. nombres TRANSCENDANTS).

1. Équations diophantiennes

Les problèmes de théorie des nombres conduisant à résoudre des équations de degré $\geqslant 2$ ont progressivement montré la nécessité d'étudier les propriétés arithmétiques des nombres algébriques et de bâtir ainsi une extension de l'arithmétique élémentaire. Le premier de ces problèmes est probablement celui qu'Euler a improprement attribué à Pell : il s'agit de résoudre en nombre entiers x et y l'équation $x^2 - Dy^2 = \pm 1$, où D est un entier positif donné, sans facteur carré. Euler remarqua très tôt que cette équation peut encore s'écrire :

$$(x + y\sqrt{D})(x - y\sqrt{D}) = \pm 1$$

et que, par suite, si (x, y) en est une solution, on en tire une infinité d'autres (u, v) en calculant $(x + y\sqrt{D})^n = u + v\sqrt{D}$ pour tout $n \in \mathbb{N}$. L'équation $x^3 + y^3 = z^3$ a fourni à Euler une autre occasion d'exploitation arithmétique de nombres irrationnels (imaginaires cette fois) ; pour établir que cette équation n'a pas de solution non triviale en nombres entiers (c'est un cas particulier du « dernier théorème de Fermat »), Euler (1770) se fonde sur le fait, admis sans démonstration, que, si p et q sont des entiers premiers entre eux tels que $(p + q\sqrt{-3})(p - q\sqrt{-3}) = p^2 + 3q^2$ soit un cube, alors chacun des deux facteurs imaginaires $p \pm q\sqrt{-3}$ est le cube d'un nombre complexe de la même forme.

Périodes

Un autre type de nombres algébriques apparaît dans la dernière section des *Disquisitiones arithmeticae* de Gauss (1801), où se trouve élaborée la théorie de l'équation de la division du cercle en n parties égales, avec n premier impair. Si r est l'une des racines imaginaires de cette équation, les autres sont $r^2, r^3, \ldots, r^{n-1}$, et Gauss introduit certaines sommes partielles de ces racines, qu'il appelle *périodes,* et qui sont solutions d'équations de degrés inférieurs : si f est un facteur de $n - 1$ et si λ est un entier quelconque, la période (f, λ) de longueur f est, par définition, la somme :

$$(f, \lambda) = r^{\lambda} + r^{\lambda h} + r^{\lambda h^2} + \ldots + r^{\lambda h^{f-1}},$$

où h est un entier premier à n tel que $h^f \equiv 1 \pmod{n}$ mais que $h^a \not\equiv 1 \pmod{n}$ si $1 \leqslant a \leqslant f - 1$; la période ne dépend pas du choix de h vérifiant ces propriétés, et on obtient un tel h en posant $h = g^e$, où g est une racine primitive modulo n (cf. DIVISIBILITÉ) et $e = (n - 1)/f$. Il y a e périodes distinctes de longueur f, correspondant à $\lambda = 1, g, g^2, \ldots, g^{e-1}$ (sans compter $(f, 0) = f$), qui sont les racines d'une équation de degré e à coefficients entiers ; ensuite, les f racines $r^{\lambda h^a}$ qui constituent la période (f, λ) sont les racines d'une équation de degré f dont les coefficients sont des combinaisons linéaires à coefficients entiers de 1 et des e périodes de longueur f. **Gauss** établit que le produit de deux périodes de longueur f est une combinaison linéaire du type précédent : ces combinaisons forment donc un sous-anneau du corps $\mathbf{C}$ des nombres complexes (cf. ANNEAUX ET ALGÈBRES) ; de plus, si p est une période de longueur f, les autres s'expriment par des polynômes en p (de degré au plus $e - 1$) à coefficients rationnels. Lorsque $e = 2$, les deux périodes de longueur $m = (n - 1)/2$ sont $(m, 1)$ et (m, g), et elles sont construites avec $h = g^2$; la première est la somme des r^a avec a résidu quadratique modulo n et la seconde la somme des r^b avec b non résidu (cf. DIVISIBILITÉ). Gauss montre que l'équation dont les racines sont ces deux périodes est $x^2 + x \pm \nu = 0$ si $n = 4\nu \pm 1$; le discriminant de cette équation est $\pm n$, dont la racine carrée est donc la valeur de la différence des deux périodes :

$$(m, 1) - (m, g) = \sum_{\lambda \bmod n} \left(\frac{\lambda}{n}\right) r^{\lambda},$$

où $\left(\frac{\lambda}{n}\right)$ est le symbole de Le Gendre. Les expressions du type :

$$\sum_{\lambda \bmod n} \left(\frac{\lambda}{n}\right) r^{\lambda} = \sum_{\lambda \bmod n} r^{\lambda^2}$$

(puisque $\sum_{\lambda \bmod n} r^{\lambda} = 0$) et leurs généralisations sont appelées *sommes de Gauss* ; elles jouent un grand rôle en théorie des nombres, et Gauss lui-même en tira deux démonstrations de la loi de

réciprocité quadratique (la quatrième, 1808, et la sixième, 1818). Si l'on précise la racine r choisie, par exemple $r = e^{2i\pi/n}$, il convient de préciser aussi celle des deux racines carrées de $\pm n$ qui donne la valeur de la somme de Gauss, et ce problème arrêta Gauss pendant longtemps. En notant $z = 0$ et $z' = 0$ les deux équations de degré m dont les racines sont respectivement les r^a (a résidu quadratique modulo n) et les r^b (b non résidu), on a :

$$\frac{x^n - 1}{x - 1} \quad zz',$$

et Gauss en déduit que :

$$4\frac{x^n - 1}{x - 1} = Y^2 \pm nZ^2,$$

où $Y = 2\ x^m + x^{m-1} + ...$ et $Z = x^{m-1} + .$ sont des polynômes en x à coefficients entiers ; si p est un nombre premier $\not\equiv 1 \pmod n$, et si $x^n \equiv 1$ (modp) (mais $x^s \not\equiv 1$ (modp) si $1 \leqslant s \leqslant n \quad 1$), on a donc $Y^2 \equiv \pm nZ^2$ (modp) et $(\pm n/p) = 1$, ce qui donne un cas particulier de la loi de réciprocité (cf. DIVISIBILITÉ).

Lorsque $e = 3$, les périodes p, p' et p'' de longueur $m = (n - 1)/3$ sont $(m, 1)$, (m, g) et (m, g^2), et ce sont les racines d'une équation du troisième degré $x^3 + x^2 \quad mx - (a^2 - bc) = 0$, où a, b et c sont des entiers tels que $pp' = bp + cp' + ap''$; ces entiers a, b, c sont aussi les nombres de solutions (x, y) modulo n pour les congruences $x^3 + 1 \equiv \lambda y^3 \pmod n$, avec $\lambda = 1$, g ou g^2, et Gauss montre que $4n = (6a \quad 3b - 3c - 2)^2 + 27(b - c)^2$. Ainsi, le quadruple d'un nombre premier n de la forme $3\ m + 1$ est représenté par la forme quadratique $x^2 + 27\ y^2$ (cf. formes QUADRATIQUES) ; comme il est facile de voir qu'une telle représentation est unique, elle donne, inversement, un moyen de déterminer les entiers a, b et c.

Lien avec les fonctions elliptiques

La théorie des fonctions elliptiques (cf. FONCTIONS ANALYTIQUES - Fonctions elliptiques et modulaire) est une autre voie par laquelle les nombres algébriques sont intervenus en mathématiques : si $\wp$ est une fonction elliptique de Weierstrass, on sait en général exprimer $\wp(nu)$ par une fonction rationnelle de $\wp(u)$ et de $\wp'(u)$ lorsque n est entier ; mais, pour certains modules, dits « singuliers », on a encore une telle expression pour $n = a \pm i\sqrt{b}$, avec a et b entiers convenables ($b \geqslant 0$). Lorsque cela se produit, on dit que la fonction elliptique admet de la multiplication complexe ; Gauss a rencontré cette situation dès la fin du XVIII[e] siècle, à propos de la fonction elliptique $x = \mathrm{sl}\, u$ (« sinus lemniscatique ») qui inverse l'intégrale :

$$u = \int_0^x \frac{dt}{\sqrt{1 - t^4}}$$

(son module J vaut 1) ; comme sl (iu) est égal à i sl u, la fonction sl admet de la multiplication complexe par tous les *entiers de Gauss* $m + ni$, où m, $n \in Z$. Abel (1828) a utilisé cette multiplication complexe pour établir que l'équation algébrique dont les racines sont les nombres $\mathrm{sl}(\omega/p)$, où p est un nombre premier de la forme $4k + 1$ et ω est l'une quelconque des périodes de la fonction sl, est résoluble par radicaux ; dans ce cas, $p = m^2 + n^2$ est le produit de deux entiers de Gauss $m \pm in$ (cf. la partie A ci-dessus - Théorie analytique des nombres), et Abel montre que la méthode de Gauss pour la division du cercle s'applique à l'équation dont les racines sont :

$$\mathrm{sl}\left(\frac{\omega}{m + in}\right).$$

Réciprocité biquadratique

Les mêmes entiers de Gauss donnent le cadre où l'on peut étudier la loi de réciprocité *biquadratique*, qui relie la résolubilité des deux congruences $x^4 \equiv p$ (mod q) et $x^4 \equiv q$ (mod p), où p et q sont des nombres premiers. Dans un article de 1832, Gauss développe l'arithmétique de ces entiers généralisés, qui repose sur un algorithme de division analogue a celui d'Euclide pour les entiers ordinaires : si a et b sont des entiers de Gauss, avec $b \neq 0$, il existe des entiers de Gauss q (« quotient ») et r (« reste ») tels que a = $bq + r$ et que, de plus, N(r) < N(b), où N désigne la *norme*, définie par :

$$N(m + in) = (m + in)(m - in) = m^2 + n^2$$

(cf. DIVISIBILITÉ). Lorsque le reste r est nul, on dit que b divise a ; les *unités,* entiers de Gauss qui divisent 1, sont les quatre nombres ± 1 et ± i (racines quatrièmes de 1), et deux entiers de Gauss sont dits associés si l'un est le produit de l'autre par une unité. Exactement comme pour les entiers ordinaires, on établit l'existence du plus grand commun diviseur (P.G.C.D.) de deux entiers de Gauss, avec une identité de Bezout ; on appelle nombre premier de Gauss un entier de Gauss qui n'est pas une unité et qui n'est divisible que par ses associés et par les unités, et on voit qu'un tel nombre ne **peut** diviser un produit de facteurs sans diviser l'un d'eux. Il en résulte que tout entier de Gauss se décompose d'une manière essentiellement unique en produit de facteurs premiers de Gauss (« essentiellement » veut dire ici « à l'ordre près des facteurs et à une multiplication près par une unité »). Les nombres premiers ordinaires se classent en trois catégories : $2 = i(1 - i)^2$, qui est associé au carré du nombre premier de Gauss $1 - i$ et se trouve seul dans sa catégorie ; les nombres premiers $\equiv 3$ (mod 4), qui restent premiers dans les entiers de Gauss ; enfin, les nombres premiers $p \equiv 1$ (mod 4), qui s'écrivent comme somme de deux carrés $p = m^2 + n^2 = (m + in)(m - in)$ et se décomposent donc, dans les entiers de Gauss, en produit de deux facteurs $m \pm in$, tous les deux premiers. Gauss introduit un symbole :

$$\left(\frac{q}{p}\right)_4$$

pour la réciprocité biquadratique analogue à celui de Le Gendre pour la réciprocité quadratique : p étant un nombre entier de Gauss et q un entier de Gauss non divisible par p,

$$\left(\frac{q}{p}\right)_4$$

est l'unité $i^k \equiv q^{(N(p)-1)/4}$ (mod p) ($0 \leqslant k \leqslant 3$), **et** il **vaut** 1 exactement dans le cas où la congruence $x^4 \equiv q$ (mod p) a une solution. La loi de réciprocité énoncée par Gauss (et démontrée par Jacobi et Eisenstein) s'énonce alors : si p et q sont des nombres premiers de Gauss non associés à $1 - i$, on a :

$$\left(\frac{p}{q}\right)_4 = (-1)^{(N(p)-1)(N(q)-1)/16}\left(\frac{q}{p}\right)_4.$$

La démonstration de Jacobi (1836) utilise des « sommes de Jacobi », intimement liées à certaines sommes de Gauss ; Eisenstein (1845) a donné d'autres démonstrations fondées sur la multiplication complexe de la fonction elliptique sl (dont l'une est inspirée par une démonstration qu'il avait trouvée pour la loi de réciprocité quadratique, et qui utilisait la formule qui donne sin nx en fonction de sin x). Jacobi et Eisenstein ont étudié de même la loi de réciprocité cubique, qui donne des rensei-

gnements sur la résolubilité d'une congruence $x^3 \equiv m \pmod{p}$ (p nombre premier ordinaire, $m \in \mathbf{Z}$) ; il faut remplacer i par la racine cubique imaginaire :

$$j = \frac{-1 + \sqrt{-3}}{2}$$

de 1, et se placer dans le cadre des nombres de la forme $m + nj$ avec $m, n \in \mathbf{Z}$. Ces nouveaux nombres ont encore des propriétés arithmétiques analogues à celles des entiers ordinaires, car ils admettent un algorithme de division euclidienne relatif à la norme $N(m + nj) = (m + nj)(m + nj^2) = m^2 - mn + n^2$; il y a six unités ± 1, $\pm j$ et $\pm j^2$ qui sont les valeurs possibles du symbole permettant d'exprimer la loi de réciprocité cubique.

« Dernier théorème de Fermat »

Avant les travaux de Kummer (cf. infra), deux nouveaux cas du « dernier théorème de Fermat », appelé aussi « grand théorème de Fermat », ont été établis par Dirichlet, Legendre et Lamé ; il s'agit de l'impossibilité de résoudre en nombres entiers non triviaux l'équation $x^n + y^n = z^n$ pour $n = 5$ (Dirichlet, Le Gendre, 1825) et pour $n = 7$ (Dirichlet, Lamé, 1839). Sophie Germain avait montré que, dans ces deux cas, l'exposant n divise nécessairement l'un des nombres x, y ou z ; pour le cas $n = 5$, Dirichlet utilise les nombres algébriques de la forme $m + n\sqrt{5}$, avec $m, n \in \mathbf{Z}$, et il établit que, si m et n sont premiers entre eux dont l'un est pair, si n est multiple de 5 et si $m^2 - 5n^2 = (m + n\sqrt{5})(m - n\sqrt{5})$ est une puissance cinquième, alors $m + n\sqrt{5}$ est la puissance cinquième d'un nombre algébrique de la même forme. Pour le cas $n = 7$, on utilise une propriété analogue des nombres algébriques de la forme $m + n\sqrt{-7}$ avec $m, n \in \mathbf{Z}$.

En 1847, Lamé crut avoir une démonstration du dernier théorème de Fermat pour tout exposant $n \geqslant 3$: il décomposait $x^n + y^n$ en n facteurs $x + r^a y$, avec $0 \leqslant a \leqslant n - 1$ et r racine n-ième de 1, et admettait que, dans le cas où x et y sont premiers entre eux, chacun de ces facteurs devait être une puissance n-ième s'il en est ainsi de leur produit $x^n + y^n = z^n$. Comme Liouville le reconnut aussitôt, la décomposition unique en facteurs premiers pour les nombres de la forme $a_0 + a_1 r + \dots + a_{n-1} r^{n-1}$ ($a_i \in \mathbf{Z}$), que Lamé admettait implicitement, n'était pas forcément justifiée. Il se trouve que Kummer, qui étudiait ces nombres depuis 1843, avait publié dès 1844 un article dans lequel il montrait que la décomposition unique en facteurs premiers n'a pas lieu pour $n = 23$.

L'intérêt de Kummer pour les « entiers cyclotomiques » du type ci-dessus provenait sans doute autant de son désir de généraliser les lois de réciprocité connues que d'efforts pour démontrer le dernier théorème de Fermat. Cependant, sa théorie lui fournit une démonstration dudit théorème pour toute une classe d'exposants n premiers, qu'il appela réguliers ; le plus petit nombre premier irrégulier est 37 (malheureusement, on ne sait pas s'il existe une infinité de nombres premiers réguliers). Rappelons que le grand théorème de Fermat a enfin été démontré en 1993.

2. Les « nombres idéaux » de Kummer

Entiers cyclotomiques

Considérons, avec Kummer, un nombre premier impair λ et une racine λ-ième imaginaire α de 1 ; ainsi :

$$\frac{\alpha^\lambda - 1}{\alpha - 1} = \alpha^{\lambda-1} + \alpha^{\lambda-2} + \dots + \alpha + 1 = 0.$$

L'équation de degré $\lambda - 1$ précédente est irréductible sur le corps Q des nombres rationnels, donc les nombres 1, a, α^2, ... $\alpha^{\wedge\ 2}$ sont linéairement indépendants sur Q ; les entiers cyclotomiques correspondant à A sont les nombres de la forme :

$$f(\alpha) = a_0 + a_1\alpha + \ldots + a_{\lambda-2}\alpha^{\lambda-2},$$

où a_0, a,. . . . . $a_{\lambda-2} \in$ ***Z,*** c'est-à-dire les éléments du sous-anneau $\mathbf{Z}[\alpha]$ de C engendré par a. Il résulte des remarques précédentes que l'écriture d'un entier cyclotomique sous la forme $f(\alpha)$, polynôme en a de degré $\leqslant \lambda$ 2 à coefficients entiers, est unique (il n'en serait pas de même avec des polynômes de degré $\leqslant \lambda$ 1 qu'il est parfois utile d'introduire). L'anneau $\mathbf{Z}[\alpha]$ des entiers cyclotomiques ne dépend pas du choix de a parmi les racines A-ièmes imaginaires de 1, puisque toutes ces racines sont les puissances de l'une d'elles ; à l'entier cyclotomique f(a), on associe ses *conjugués* $f(\alpha^2), f(\alpha^3)$, ... $f(\alpha^{\lambda\ 1})$ et sa *norme* $\mathrm{N}f(\alpha) = f(\alpha)f(\alpha^2)\ldots f(\alpha^{\wedge\ 1})$, qui est un entier ordinaire car elle ne change pas lorsqu'on remplace α par une de ses puissances. Comme α^j et $\alpha^{\lambda\ j}$ sont complexes conjugués, $f(\alpha^j)f(\alpha^{\lambda\ j})$ est réel positif, et $\mathrm{N}f(\alpha)$ est donc un entier positif ; on vérifie facilement que la norme est multiplicative : $\mathrm{N}(f(\alpha)g(\alpha)) = \mathrm{N}f(\alpha)\cdot \mathrm{N}g(\alpha)$. On dit qu'un entier cyclotomique $f(\alpha)$ en divise un autre /z(a) s'il existe un entier cyclotomique g(a) tel que $h(\alpha) = f(\alpha)g(\alpha)$; on dit que $h_1(\alpha)$ et $h_2(\alpha)$ sont congrus modulo $f(\alpha)$, et on écrit h,(a) $\equiv h_2(\alpha) \pmod{f(\alpha)}$, si $h_1(\alpha) - h_2(\alpha)$ est divisible par $f(\alpha)$. Kummer appelle *unités* les entiers cyclotomiques dont la norme est 1, c'est-à-dire ceux qui divisent 1 ; par exemple les α^j sont des unités. de même que les quotients :

$$\frac{1-\alpha^j}{1\text{ - a}} = 1 + \text{a} + \alpha^2 + + \alpha^{j-1}$$

(avec $1 \leqslant j \leqslant$ A-1), car $\mathrm{N}(1 - \alpha^j) = (1 - \alpha^j)(1 - \alpha^{2j}) \ldots (1 - \alpha^{(\wedge\ 1)j})$ est la valeur, pour $x = 1$, du polynôme $(x - \alpha^j)$ $(x - \alpha^{(\lambda\ 1)j}) = 1 + x + x^2 + \ldots + x^{\lambda-1}$, ce qui donne $\mathrm{N}(1 - \alpha^j) = \lambda$ indépendamment dej. Des calculs précédents, on titre aussi :

$$\text{A} = (1 - \alpha)^{\lambda-1}\frac{1-\alpha^2}{1-\alpha}\,\frac{1-\alpha^3}{1-\alpha}\cdots\frac{1-\alpha^{\lambda-1}}{1\text{ - a}};$$

ainsi λ se décompose, dans l'anneau des entiers cyclotomiques, en le produit de $(1\quad\alpha)^{\lambda-1}$ par une unité.

On dit qu'un entier cyclotomique $h(\alpha)$ est *premier* s'il n'est pas une unité et s'il ne peut diviser un produit $f(\alpha)g(\alpha)$ sans diviser l'un des facteurs f(a) ou g(a) ; si h(a) est premier, il en est de même de ses conjugués $h(\alpha^j)$ et des produits de ces conjugués par des unités. Par exemple, 1 a est premier : en effet, si 1 a divise un entier rationnel m, A $=$ N(1 $-$ a) divise $\mathrm{N}m = m^{\lambda-1}$, c'est-à-dire que A divise m (puisque A est premier) et, inversement, les multiples de λ sont divisibles par 1 a ; si 1 - a divise $f(\alpha)g(\alpha)$, c'est-àdire $f(\alpha)g(\alpha) \equiv 0 \pmod{1-\alpha}$, comme $\alpha \equiv 1$ (mod(l-a)), on a $f(1)g(1) \equiv 0$ (mod (1 ***a)),*** donc f(1) ou g(1) est divisible par λ et, par suite, f(a) $\equiv 0$ ou ***g(a)*** $\equiv$ ***0*** (mod (1 $-$ ***a)).*** Soit /z(a) un entier cyclotomique de norme A ; le nombre premier cyclotomique 1 ***a*** divise $\mathrm{N}(h(\alpha))$, donc divise l'un de ses facteurs $h(\alpha^j)$ $(1 \leqslant j \leqslant$ A $- 1)$, et, comme $\mathrm{N}(h(\alpha^j)) = \mathrm{N}(h(\alpha)) = \lambda = N($ 1 $-$ a), le quotient est une unité : h(a) est donc encore premier et c'est le produit d'un des conjugués de 1 $-$ ***a*** par une unité.

Considérons maintenant un entier cyclotomique $h(\alpha)$ dont la norme soit un nombre premier $q \neq \lambda$; ainsi ***q*** = $\mathrm{N}(h(\alpha)) \equiv h(1)^{\lambda\ 1}$ (mod (1 -a)), d'où, d'après ce qui précède $q \equiv h(1)^{\lambda-1} \equiv 1$

(mod λ) (théorème de Fermat ; cf. DIVISIBILITÉ). On voit comme ci-dessus que les entiers rationnels divisibles par $h(\alpha)$ sont les multiples de q ; pour continuer le raisonnement et montrer que $h(\alpha)$ est premier, on va prouver qu'il existe un entier rationnel u tel que a $\equiv u$ (mod $h(a)$), et on lui fera jouer le rôle que jouait 1 pour 1 $-$ a. Posons $i = (q \quad 1)/\lambda$ et k = y', où y est une racine primitive modulo q ; la congruence $x^\lambda \equiv 1$ (mod q) a pour solutions x $\equiv 1, k, k^2, \ldots . k^{\lambda-1}$ (mod q), donc $x^{\lambda \; 1} + x^{\lambda-2} + \ldots + x + 1 \equiv (x \quad k)(x \quad k^2)\,(x \quad k^{\lambda-1})$ (mod q). On en déduit facilement que $h(k)h(k^2) \; h(k^{\lambda-1}) \equiv 0$ (mod q), puisque $N(h(\alpha)) = q$; par suite, q divise l'un des nombres $h(k^j)$ et le polynôme h(x) est divisible modulo q par x $- k^j$. On prend $u = k^j$, de sorte que $h(\alpha^m)$ est divisible modulo q par $\alpha^m \quad u$ pour $1 \leqslant$ m $\leqslant \lambda - 1$, et $(a \quad u)h(\alpha^2)h(\alpha^3) \,.\, h(\alpha^{\lambda-1})$ est divisible modulo q par :

$$N(a-u) = \frac{1-u''}{1-u},$$

lui-même divisible par $q = h(\alpha)h(\alpha^2) \; h(\alpha^{\lambda-1})$; cela prouve bien que $a - u$ est divisible par h(a). Il en résulte alors que, si $h(\alpha)$ divise $f(\alpha)g(\alpha)$, il divise $f(u)g(u)$ qui est donc un multiple de q ; ainsi q divise f(u) ou g(u), ce qui signifie que $h(a)$ divise $f(a)$ ou $g(a)$: on voit donc que h(a) est premier. Si h,(a) est un autre entier cyclotomique de norme q, $h(a)$, qui divise q, divise l'un des conjugués de h,(a), et le quotient est de norme 1 ; ainsi, $h,(a)$, qui est aussi premier, est le produit d'un des conjugués de h(a) par une unité. Les différents facteurs premiers $h(a), h(\alpha^2), \ldots . h(\alpha^{\lambda-1})$ trouvés pour q dans les entiers cyclotomiques sont essentiellement distincts, c'est-à-dire qu'aucun d'eux n'en divise un autre ; si en effet $h(\alpha^j)$ divisait $h(\alpha^m)$, de $\alpha^j \equiv u$ (mod $h(\alpha^j)$) on tirerait $\alpha^j \equiv u \equiv \alpha^m$ (mod $h(\alpha^m)$), donc $q = N(h(\alpha^m))$ diviserait $N(\alpha^j \quad \alpha^m) = N(1 \quad \alpha^{m-j})$, ce qui exige $m \equiv j$ (mod λ) (sinon $N(1 - \alpha^{m-j}) = \lambda$).

Kummer a fait des calculs systématiques de nombres premiers cyclotomiques h(a) de norme un nombre premier rationnel q, et des entiers u correspondants. en prenant de petites valeurs de A. Par exemple, pour $\lambda = 5$, $a + 2$ a pour norme 11, $a - 2$ a pour norme 31, etc. ; pour $\lambda = 7, - \alpha^4 + \alpha^2 + 1$ a pour norme 29 et divise a + 13, a + 2 a pour norme 43. etc. Ses tables donnent, pour A $\leqslant$ 19, les facteurs premiers cyclotomiques de tous les nombres premiers $q \equiv 1$ (mod A) inférieurs à 1 000. Pour A = 23, on voit que 47 n'est la norme d'aucun entier cyclotomique, et n'admet donc aucun diviseur premier cyclotomique ; cependant :

$$N(a-4) = \frac{4^{23}-1}{4-1} \equiv 0 \pmod{47},$$

et on calcule que $N(1 \quad a + a^{-2}) = 47 \cdot 139$ et $N(\alpha^{10} + \alpha^{-10} + \alpha^8 + a^{8} + \alpha^7 + \alpha^{-7}) = 47^2$, 47 étant le produit de la moitié des conjugués de ce dernier nombre. On en déduit deux décompositions distinctes de 47 . 139 en produits d'entiers cyclotomiques irréductibles, mais non premiers. L'idée des « facteurs premiers idéaux » de Kummer consiste à associer à $a - 4$ un tel facteur : par définition, un entier cyclotomique $f(\alpha)$ est divisible par ce facteur sif(4) est divisible par 47, mais le facteur lui-même n'existe pas en tant que nombre.

Si $h(a)$ est un nombre premier cyclotomique quelconque, il divise $N(h(\alpha))$, donc l'un des facteurs premiers rationnels y de cet entier ; montrons que les entiers rationnels divisibles par $h(a)$ sont exactement les multiples de q : si $h(a)$ divise

l'entier m, il divise le P.G.C.D. (m, q) qui vaut 1 (exclu car $h(\alpha)$ n'est pas une unité) ou q, et par suite q divise m. Lorsque $q \neq$ A, on a $q^{\lambda\ 1} \equiv 1$ (mod A) par le petit théorème de Fermat ; soit f le plus petit entier $\geqslant 1$ tel que $q^f \equiv 1$ (mod A). On sait quefest un diviseur de A $- 1$; le cas $f = 1$ a été traité ci-dessus, et nous allons examiner le cas général en cherchant des entiers cyclotomiques congrus à des entiers rationnels modulo $h(\alpha)$. D'après le théorème de Fermat, on a :

$$x^q - x \equiv (x-1)(x-2)\ldots(x-q) \pmod q;$$

en prenant $x = f(\alpha)$ entier cyclotomique tel que $f(\alpha^q) = f(\alpha)$, donc f(o) $= f(\alpha^q) \equiv f(\alpha)^q$ (modq), on voit que h(o) divise l'un des facteurs $f(\alpha) - u$, c'est-à-dire que $f(\alpha) \equiv u$ (mod $h(\alpha)$). La transformation $\alpha \mapsto \alpha^q$ laisse invariantes les e = (A 1)/f périodes $\eta_0, \eta_1, \ldots.\ \eta_{e-1}$ de longueur f, et il existe donc des entiers rationnels $u_0, u_1, \ldots.\ u_{e-1}$ tels que $\eta_i \equiv u_i$ (mod $h(\alpha)$) pour $0 \leqslant i \leqslant e-1$; autrement dit, l'équation de degré e (à coefficients entiers rationnels) dont les solutions sont les périodes de longueur f se décompose complètement modulo q en $(x-u_0)(x-u_1)$ (x $- u_{e-1}$).

Pour qu'un entier de la forme a, + $a_1 \eta_1 + + a_e\eta_e$ ($\eta_e = \eta_0$, les indices des périodes sont pris modulo e), avec a_0, a,, ... $a_e \in Z$, soit divisible par $h(\alpha)$, il faut et il suffit que $a_0 + a_1u_1 + + a_eu_e$ soit divisible par q ; chacun des entiers cyclotomiques formés des périodes η_i est congru modulo h(o) à un unique entier rationnel appartenant à $[0, q-1]$. Comme α est racine d'une équation de degré f à coefficients formés des périodes de longueur f, tout entier cyclotomique écrit d'une manière unique sous la forme $\psi(\eta) + \alpha\psi_1(\eta) + + \alpha^{f-1}\psi_{f-1}(\eta)$, où les coefficients $\psi_j(\eta)$ sont formés des périodes de longueur f, et il est donc congru modulo h(o) à un unique entier cyclotomique de la forme $b_0 + b_1\alpha + + b_{f-1}\alpha^{f-1}$, où les coefficients b_j sont des entiers rationnels de l'intervalle $[0, y-1]$; les q^f entiers cyclotomiques ainsi décrits forment donc un système de représentants des classes modulo $h(\alpha)$. Les entiers u_i au moyen desquels on peut tester la divisibilité par h(o) sont liés par des relations modulo q qui proviennent des relations entre les périodes : si $\eta_i\eta_j = a_0 + a_1\eta_1 + + a_e\eta_e$, on doit avoir $u_iu_j \equiv a_0 + a_1u_1 + + a_eu_e$ (modq).

Considérons maintenant un nombre premier rationnel $y \neq$ A quelconque, et soitfle plus petit entier $\geqslant 1$ tel que $q^f \equiv 1$ (mod λ) ; on pose encore e = (A $- 1)/f$. Il se peut que q ne soit divisible par aucun nombre premier cyclotomique, comme nous l'avons vu dans le cas où λ = 23 et q = 47 ; mais on peut encore démontrer que l'équation de degré e dont les racines sont les périodes de longueur f se décompose modulo q en e équations de degré 1, et les e entiers rationnels, racines modulo q de ces congruences, vont permettre de définir une relation de congruence dans les entiers cyclotomiques, si on les associe convenablement aux e périodes de longueur f. Pour ce faire, on considère les entiers cyclotomiques $u - \eta_i$, avec u entier rationnel tel que :

$$\prod_{i=0}^{e-1} (u - \eta_i)$$

soit divisible par q ($0 \leqslant u \leqslant q$ 1), et on forme un produit, noté $\Psi(\eta)$, de tels nombres avec des u et des η_i choisis tels que $\Psi(\eta)$ ne soit pas divisible par q, mais que, pour tout i, il existe u_i tel que (u, $\eta_i)\Psi(\eta) \equiv 0$ (mod q) ; l'entier rationnel u_i est déterminé d'une manière unique par

cette condition, car si on en avait deux, u_i et u', on aurait $(u_i - u')\Psi(\eta) \equiv 0 \pmod q$, avec $u_i - u' \not\equiv 0 \pmod q$, donc inversible modulo p, ce qui donnerait $\Psi(\eta) \equiv 0 \pmod q$. Soit $F(\eta_0, \eta_1, \ldots, \eta_{e-1}) = 0$ une relation polynomiale quelconque (à coefficients entiers) entre les périodes ; on tire de ce qui précède que $F(u_0, u_1, \ldots, u_{e-1})\Psi(\eta) \equiv 0 \pmod q$, d'où $F(u_0, u_1, \ldots, u_{e-1}) \equiv 0 \pmod q$, puisque $F(u_0, u_1, \ldots, u_{e-1})$ est un entier rationnel et que $\Psi(\eta) \not\equiv 0 \pmod q$. Kummer associe à la suite $(u_0, u_1, \ldots, u_{e-1})$ un « nombre premier idéal » divisant q, qui n'est pas du tout un nombre, mais bien une relation d'équivalence entre entiers cyclotomiques, pour laquelle chaque période η_i de longueur f est équivalente à l'entier rationnel u_i correspondant. Cela détermine entièrement la relation d'équivalence, si on lui impose d'être compatible avec la structure d'anneau des entiers cyclotomiques (addition et multiplication ; cf. **ANNEAUX ET ALGÈBRES**) ; explicitement, des entiers cyclotomiques sont équivalents si leurs produits par $\Psi(\eta)$ sont congrus modulo q. On montre encore que les q^f entiers cyclotomiques $b_0 + b_1\alpha + \ldots + b_{f-1}\alpha^{f-1}$, avec $0 \leqslant b_j \leqslant q-1$ pour $j = 0, 1, \ldots, f-1$, forment un système de représentants des classes pour la relation d'équivalence considérée, et qu'un produit d'entiers cyclotomiques ne peut être équivalent à 0 sans que l'un des facteurs le soit : autrement dit, l'anneau quotient est intègre et, comme il est fini, c'est un corps à q^f éléments.

Le groupe de Galois de l'équation dont les racines sont les e périodes de longueur f est formé des permutations circulaires de ces périodes (cf. **CORPS**) ; on peut donc transformer la suite $(u_0, u_1, \ldots, u_{e-1})$ par permutations circulaires, ce qui définit en tout e « nombres premiers idéaux » divisant q. On peut vérifier qu'ils sont tous distincts, et que ce sont les seuls possibles ; en ce sens, q se trouve décomposé, dans les entiers cyclotomiques, en $e = (\lambda - 1)/f$ facteurs premiers distincts, tous conjugués, mais pouvant être idéaux : pour qu'un entier cyclotomique soit divible par q, il faut et il suffit qu'il soit divisible par chacun des e facteurs de q, c'est-à-dire équivalent à 0 pour chacune des e relations d'équivalence correspondantes. Par exemple, le nombre $\Psi(\eta)$ qui a servi à construire $(u_0, u_1, \ldots, u_{e-1})$ est divisible par tous les facteurs de q sauf celui qui est associé à $(u_0, u_1, \ldots, u_{e-1})$. Notons que, lorsque $f = \lambda - 1$, c'est-à-dire lorsque q est une racine primitive modulo λ, $e = 1$ et q reste premier dans les entiers cyclotomiques.

On précise la définition des « nombres premiers idéaux » en spécifiant avec quelle multiplicité ils divisent tel ou tel entier cyclotomique. Soit, comme plus haut, $\Psi(\eta)$ un entier cyclotomique formé des périodes de longueur f et divisible par tous les facteurs d'un nombre premier rationnel q d'ordre f modulo λ, sauf celui qui correspond à une suite $(u_0, u_1, \ldots, u_{e-1})$ d'entiers rationnels ; on définit une *valuation* v sur l'anneau des entiers cyclotomiques en posant :

$$v(f(\alpha)) = \sup\{n \in \mathbf{N} \mid f(\alpha)\Psi(\eta)^n \equiv 0 \pmod{q^n}\}$$

(cf. algèbre **TOPOLOGIQUE**). Lorsque le facteur de q associé à $(u_0, u_1, \ldots, u_{e-1})$ est un vrai nombre premier cyclotomique $h(\alpha)$, les nombres de valuation n sont ceux qui sont divisibles par $h(\alpha)^n$ mais pas par $h(\alpha)^{n+1}$. La théorie de Kummer consiste en définitive à remplacer les nombres premiers cyclotomiques, qui n'existent pas toujours, par des valuations ; au moyen de ces valuations, on peut énoncer un critère de divisibilité pour les entiers cyclotomiques, qui joue le rôle de la décomposition en facteurs premiers lorsque celle-ci est possible.

Théorème. Soit $f(\alpha)$ et $g(a)$ deux entiers cyclotomiques. Pour que $f(\alpha)$ divise $g(a)$, il faut et il suffit que, pour toute valuation ν associée à un « nombre premier idéal », on ait $\nu(f(\alpha)) \leqslant \nu(g(a))$.

La condition est évidemment nécessaire, car les valuations ν ne prennent que des valeurs positives sur les entiers cyclotomiques ; pour voir qu'elle est suffisante, on observe que $f(\alpha) | g(\alpha)$ équivaut à $Nf(\alpha) | g(\alpha)f(\alpha^2)f(\alpha^3)\ldots f(\alpha^{\lambda-1})$, donc on peut se ramener au cas où $f(a)$ est un entier rationnel, en ajoutant $\nu(f(\alpha^2)\ldots f(\alpha^{\lambda-1}))$ à $\nu(f(\alpha))$ et à $\nu(g(a))$. On décompose alors $f'(a)$ en un produit $p_1 p_2 \; p_n$ de nombres premiers rationnels (non nécessairement distincts), et on se ramène, par récurrence sur n, à montrer que la propriété est vraie pour $f(\alpha) = q$, nombre premier rationnel, ce qu'on a déjà vu plus haut.

En résumé, un « nombre premier idéal » Q est défini par la donnée d'un nombre premier rationnel $q \neq$ A et d'une correspondance convenable entre les e périodes de longueurf(ordre de q modulo λ) et les racines modulo q de l'équation de degré e qui donne ces périodes ; il faut ajouter à cela le nombre premier exceptionnel 1 — a, unique diviseur de λ avec la multiplicité A — 1. À l'entité Q, on associe une valuation vo sur l'anneau des entiers cyclotomiques ; si $\nu_Q(f(\alpha)) \geqslant 1$, on a $f(\alpha) \equiv 0$ (mod q) (et inversement), donc Q divise $N(f(\alpha))$, qui est ainsi multiple de q : on a donc $\nu_Q(f(\alpha)) = 0$ sauf pour un nombre fini de Q qui divisent les facteurs premiers rationnels de $N(f(\alpha))$. À l'entier cyclotomique $f'(a)$, on associe le symbole :

$$\prod_Q Q^{\nu_Q(f(\alpha))} = (f(\alpha)),$$

que l'on appelle *le diviseur* de $f(a)$, et on voit, grâce au théorème précédent, que deux entiers cyclotomiques n'ont le même diviseur que si l'un est le produit de l'autre par une unité (pour une unité $e(a)$, on a $\nu_Q(e(\alpha)) = 0$ quel que soit Q). D'une manière générale, on appelle diviseur un symbole $P_1^{n_1}P_2^{n_2}\,.\,P_k^{n_k}$, où P_1, $P_2, \ldots . P_k$ sont des « nombres premiers idéaux » (que l'on peut identifier aux valuations correspondantes ν_{P_i}) et $n_1, n_2, \ldots . n_k \in N$; on peut multiplier entre eux les diviseurs (mais pas les additionner). et la multiplication est une loi commutative et associative admettant comme élément neutre le diviseur (1) des unités. Si A est un diviseur, on note encore vo(A) l'exposant du « nombre premier idéal » Q dans A ; à deux diviseurs A et B, on associe un diviseur P.G.C.D. (A, B), pour lequel :

$$\nu_Q((A, B)) = \inf(\nu_Q(A), \nu_Q(B))$$

quel que soit Q. Par exemple un « nombre premier idéal » P, facteur d'un nombre premier rationnel q, s'écrit comme le P.G.C.D. $(q, \psi(\eta))$ des diviseurs (y) et (w(n)), où $\psi(\eta)$ est un entier cyclotomique formé de périodes de longueur f(ordre de q modulo A) tel que $\nu_P(\psi(\eta)) = 1$ et $\nu_Q(\psi(\eta)) = 0$ pour les autres facteurs idéaux Q de q ; il est facile de construire un tel $\psi(\eta)$ en utilisant le nombre B'(n) introduit plus haut et ses conjugués.

Soit σ un automorphisme de conjugaison des entiers cyclotomiques, défini par une substitution $\alpha \mapsto \alpha^j$; on fait opérer σ sur les diviseurs en posant $\sigma(P_1^{n_1} P_2^{n_2} \ldots P_k^{n_k}) = \sigma(P_1)^{n_1} \sigma(P_2)^{n_2}$ $\sigma(P_k)^{n_k}$, et $\sigma(P) = (p, \sigma\psi(\eta))$ si $P = (p, \psi(\eta))$. La norme d'un diviseur A est le diviseur NA = A . o(A) ... $\sigma^{\lambda-2}(A)$, où σ est un automorphisme de conjugaison défini par a $\mapsto \alpha^\gamma$, y racine primitive modulo A ; on voit facilement que c'est le diviseur d'un entier rationnel que l'on note encore NA. L'entier NA s'interprète

encore comme le nombre de classes d'entiers cyclotomiques pour la congruence modulo A, en définissant f(a) $\equiv g(\alpha)$ (mod A) par la condition : A divise $(f(\alpha) - g(\alpha))$; pour le voir, on se ramène, grâce à une généralisation du théorème des restes chinois (cf. équations DIOPHANTIENNES), au cas où A = P^n est une puissance d'un diviseur premier P $= (p, \psi(\eta))$, et on établit, par récurrence sur n, que tout entier cyclotomique est congru modulo P^n à un nombre de la forme :

$$a_0 + a_1 \psi(\eta) + \ldots + a_{n-1} \psi(\eta)^{n-1},$$

où les a, sont des entiers cyclotomiques déterminés d'une manière unique modulo P. Ainsi le nombre de classes modulo P^n est la puissance n-ième du nombre de classes modulo P, et on est ramené à $n = 1$, soit A = P; alors P a seulement e conjugués distincts $(p, \sigma^j \psi(\eta))$, avec $0 \leqslant j \leqslant e - 1$, qui sont répétés chacun f fois, et le produit de ces conjugués est p : on a donc NP $= p^f$ qui est bien le nombre de classes modulo P. Le cas où P = 1 α est à part ; il y a A $=$ N(1 α) classes modulo (1 α), représentées par 0, 1, ... $\lambda - 1$.

Classes de diviseurs

Les diviseurs premiers ont été définis comme facteurs de nombres premiers rationnels ; par suite, tout diviseur A divise un nombre véritable : il existe un diviseur C tel que AC soit le diviseur d'un entier cyclotomique. On appelle *principaux* les diviseurs d'entiers cyclotomiques. Considérons des diviseurs A, B, C tels que AC et BC soient principaux, et soit D un diviseur tel que AD soit principal ; on a donc AC = $(f(\alpha))$, BC = $(g(\alpha))$ et AD = $(h(\alpha))$, et on voit, en se servant du théorème du chapitre précédent, que $f(\alpha)$ divise $g(\alpha)h(\alpha)$ dont le diviseur est ACBD. Le quotient est un entier cylotomique de diviseur BD, qui est donc encore principal ; si, donc, pour un diviseur C, AC et BC sont tous deux principaux, alors AD et BD sont principaux en même temps pour un diviseur quelconque D. Selon Kummer, on dit que les diviseurs A et B sont équivalents s'il existe un diviseur C tel que AC et BC soient principaux ; les diviseurs principaux sont ceux qui sont équivalents au diviseur (1). La relation d'équivalence est compatible avec la multiplication des diviseurs : si A, (resp. A_2) est équivalent à B, (resp. B_2), A_1A_2 est équivalent à B_1B_2, car si A_1A_2C est principal, il en est de même de $B_1A_2C = A_2B_1C$ (A, équivalent à B,), donc aussi de B_2B_1C (A_2 équivalent à B_2). Pour chaque diviseur A, il existe un diviseur C tel que AC soit équivalent à l'élément neutre (1) de la multiplication ; par suite, la multiplication des diviseurs induit, sur l'ensemble des classes de diviseurs, une structure de *groupe commutatif*, dont l'élément neutre est la classe des diviseurs principaux.

Un théorème fondamental de Kummer affirme que le groupe des classes de diviseurs est *fini*. Sa démonstration repose sur deux lemmes.

Lemme 1. Il n'y a qu'un nombre fini de diviseurs de norme inférieure à un entier M fixé.

On a en effet N(($1 - \alpha)^{n_0}$ $P_1^{n_1}$ $P_2^{n_2}$ $P_k^{n_k}$) $= \lambda^{n_0} p_1^{f_1 n_1} p_2^{f_2 n_2} \ldots p_k^{f_k n_k} \leqslant$ M, donc les p_i et les n_i ne peuvent prendre qu'un nombre fini de valeurs ; chaque p_i n'a qu'un nombre fini e_i de facteurs premiers idéaux, donc les P_i ne prennent qu'un nombre fini de valeurs, ce qui démontre le lemme.

Lemme 2. Soit $\mu = (A - 1)/2$; pour tout diviseur A, il existe un diviseur C de norme $\leqslant \lambda^{\mu}$ tel que AC soit principal.

Soit en effet c le plus petit entier tel que $(c+1)^{\lambda-1} \geqslant$ NA + 1 ; les entiers cyclotomiques $a_1\alpha + a_2\alpha^2 + + ah -$ la" ^{1} tels que $0 \leqslant a_i \leqslant c$ (*i* = 1, ... A — 1) sont en nombre supérieur à NA, donc il y en a deux qui sont congrus modulo A et on prend pour AC le diviseur de la différence de ces deux nombres, qui s'écrit $b_1\alpha + b_2\alpha^2 + \quad + b_{\lambda-1}\alpha^{\lambda-1} = f(\alpha)$ avec $|b_i \leqslant c$. On a :

$$\prod_{i=1}^{\mu} f(\alpha^i) f(\alpha^{-i}) \leqslant \left(\frac{1}{\mu}\sum (f(\alpha^i) f(\alpha^{-i})\right)^{\mu}$$

(inégalité de la moyenne géométrique, applicable aux nombres réels $f(\alpha')f(\alpha\ ')$), et on calcule facilement que :

$$\sum_{i=1}^{h-l} (f(\alpha^i) f(\alpha^{-i})) = \lambda \sum b_i^2 - \left(\sum b_i\right)^2 ;$$

donc :

$$\sum_{i=1}^{\lambda-1} f(\alpha^i) f(\alpha^{-i})) \leqslant \lambda(\lambda-1)c^2$$

et :

$$\frac{1}{\mu}\sum_{i=1}^{\mu} f(\alpha^i) f(\alpha^{-i})) \leqslant \lambda c^2,$$

d'où $N(f(\alpha)) \leqslant \lambda^{\mu} c^{\lambda\ 1} \leqslant \lambda^{\mu}$NA d'après le choix de c, ce qui signifie que NC $\leqslant \lambda^{\mu}$.

En appliquant ces lemmes, on voit qu'il existe un nombre fini de diviseurs C,, C_2, ... C,,, tels que, pour tout diviseur A, l'un des AC, $(1 \leqslant j \leqslant m)$ soit principal ; donc il y a au plus m classes de diviseurs distinctes.

3. Unités

Dans une série dc courtes notes, Dirichlet (1841-1846) a étudié les unités dans des anneaux de nombres algébriques de la forme **Z**[θ], où θ vérifie une équation irréductible $x^n + a_1x^{n-1} + + a_{,,} = 0$ à coefficients *a,* entiers rationnels ; si les racines de cette équation sont $\theta, \theta_1, \ldots . \theta_{n-1}$, les conjugués d'un élément f'(B) de **Z**[θ] sont $f(\theta_1), \ldots . f(\theta_{n-1})$, et sa norme est le produit $N(f(\theta)) = f(\theta)f(\theta_1) \ldots f(\theta_{n-1})$. Les unités de **Z**[θ] sont les éléments $f(\theta)$ de norme ± 1 (la norme est toujours un entier rationnel, mais elle peut être négative dans ce cas général) ; parmi les unités, les racines de 1 qui appartiennent à **Z**[θ] sont caractérisées par $|f(\theta)| = \quad |f(\theta_1)| = \quad = |f(\theta_{n-1})| = \quad 1$. En effet, si $f(\theta)$ vérifie ces conditions, il en est de même de ses puissances $f(\theta)^k$, $k \in N$, dont tous les conjugués restent donc bornés. Or l'ensemble des éléments de **Z**[θ] dont les conjugués sont tous majorés par une constante M est fini, car ces éléments sont les racines d'un nombre fini d'équations de degré n (leurs coefficients sont les fonctions symétriques élémentaires des conjugués, donc ce sont des entiers rationnels majorés en fonction de M et de n) ; il n'y a donc qu'un nombre fini de puissances $f(\theta)^k$ distinctes, et $f(\theta)^l = 1$ pour 1 convenable. Les racines de 1 appartenant à **Z**[θ] forment un groupe fini cyclique pour la multiplication ; la finitude provient du résultat précédent, et le caractère cyclique du fait que, pour tout l, il y a au plus l solutions de l'équation $x^l = 1$ dans C, donc dans **Z**[θ] (cf. GROUPES - Généralités). L'énoncé fondamental de Dirichlet est le suivant :

Théorème. Soit r_1 le nombre de racines réelles de l'équation en θ, et 2 r_2 le nombre de ses racines imaginaires (de sorte que *Y, + 2* r_2 = n). Il existe Y = Y, + r_2 1 unités fondamentales $e_1(\theta), e_2(\theta), \ldots .$ r,.(O) telles que toute unité s'écrive, d'une manière unique, sous la forme $\omega e_1(\theta)^{n_1}\ e_2(\theta)^{n_2}\ e_r(\theta)^{n_r}$, où ω est une

racine de 1 et où les exposants n_i appartiennent à **Z**.

Autrement dit, le groupe multiplicatif des unités est le produit du groupe des racines de 1 par un groupe isomorphe à $\mathbf{Z}^r$.

On démontre le théorème en utilisant le plongement logarithmique ainsi défini : on indexe les racines de l'équation en θ de manière que $\theta_1, \theta_2, \ldots, \theta_{r_1}$ soient réelles et que θ_{j+r_2} soit complexe conjugué de θ_j pour $r_1 + 1 \leqslant j \leqslant r_1 + r_2$; on note alors $Le(\theta)$ le vecteur $(\ln |e(\theta_i)|)$ de $\mathbf{R}^{r_1+r_2}$ $(1 \leqslant i \leqslant r_1 + r_2)$. Ainsi $e(\theta) \mapsto Le(\theta)$ est un homomorphisme du groupe des unités dans $\mathbf{R}^{r_1+r_2}$, et son noyau est le sous-groupe formé des racines de 1 ; l'image est un sous-groupe de $\mathbf{R}^{r_1+r_2}$, et on aura démontré le théorème en prouvant que cette image est un groupe libre de rang r. Or l'image de L est discrète, car si $Le(\theta)$ reste borné, tous les conjugués de $e(\theta)$ sont majorés en valeur absolue par une constante et $e(\theta)$ ne peut prendre qu'un nombre fini de valeurs ; on sait qu'un sous-groupe discret de $\mathbf{R}^s$ est libre de rang $\leqslant s$ (cf. algèbre TOPOLOGIQUE). En écrivant que la norme de $e(\theta)$ est $\pm$ 1, on voit de plus que l'image de L est contenue dans l'hyperplan d'équation :

$$\sum_{i=1}^{r_1} x_i + 2\sum_{j=1}^{r_2} x_{r_1+j} = 0,$$

ce qui majore son rang par $r_1 + r_2 - 1 = r$; cet hyperplan se projette isomorphiquement sur $\mathbf{R}^r$, et on note $\tilde{L}$ la composée de L avec la projection. Il reste à voir que l'image de $\tilde{L}$ est de rang r, c'est-à-dire que l'orthogonal de cette image dans le dual de $\mathbf{R}^r$ est 0 (cf. algèbre LINÉAIRE ET MULTILINÉAIRE) ; on obtient ce résultat grâce au théorème de Minkowski (cf. approximations DIOPHANTIENNES), qui permet de prouver l'existence d'un élément non nul $f(\theta)$ de $\mathbf{Z}[\theta]$ vérifiant les inégalités $|f(\theta_i)| \leqslant \kappa_i$ $(1 \leqslant i \leqslant n)$, avec des nombres réels positifs $\kappa_1, \kappa_2, \ldots, \kappa_n$ tels que $\kappa_{j+r_2} = \kappa_j$ pour $r_1 + 1 \leqslant j \leqslant r_1 + r_2$ et que le produit $\kappa_1 \kappa_2 \ldots \kappa_n$ soit assez grand. Si $f(\theta) = a_0 + a_1\theta + \ldots + a_{n-1}\theta^{n-1}$, ces inégalités s'interprètent comme un système d'inégalités définies par des jauges de Minkowski sur l'espace $\mathbf{R}^{r_1} \times \mathbf{C}^{r_2} \simeq \mathbf{R}^n$ des vecteurs $(a_0, a_1, \ldots, a_{n-1})$; la matrice des jauges est $(\theta_i^j)_{1 \leqslant i \leqslant n, 0 \leqslant j \leqslant n-1}$, et la condition d'existence de $f(\theta)$ est que $\kappa_1\kappa_2 \ldots \kappa_n$ soit plus grand que la valeur absolue Δ du déterminant de cette matrice. L'élément $f(\theta)$ ainsi obtenu est de norme $f(\theta_1)f(\theta_2) \ldots f(\theta_n)$ au moins 1 en valeur absolue (entier non nul), ce qui donne $|f(\theta_i)| \geqslant \kappa_i/K$ pour $i = 1, 2, \ldots, n$, en posant $K = \kappa_1\kappa_2 \ldots \kappa_n$. Soit φ une forme linéaire non nulle sur $\mathbf{R}^r$ et x le vecteur de coordonnées $(\ln \kappa_1, \ln \kappa_2, \ldots, \ln \kappa_r)$; les inégalités obtenues donnent $|\varphi(x) - \varphi(\tilde{L}(f(\theta)))| \leqslant \|\varphi\| \ln K$ (où $\|\varphi\|$ est la somme des valeurs absolues des coefficients de φ). On prend $M > \|\varphi\| \ln K$ fixé, et, pour chaque entier h, on choisit $\kappa_1, \kappa_2, \ldots, \kappa_r$ de manière que $\varphi(x) = 2Mh$; on peut alors trouver κ_{r+1} assez petit pour que $\|\varphi\| \ln (\kappa_1\kappa_2 \ldots \kappa_n) < M$, et on a un élément $f_h(\theta)$ correspondant dans $\mathbf{Z}[\theta]$ qui vérifie $(2h - 1)M < \varphi(\tilde{L}(f_h(\theta))) < (2h + 1)M$, de sorte que la suite $(\varphi \circ \tilde{L}(f_h(\theta)))_h$ est strictement croissante. Par ailleurs, les normes $Nf_h(\theta)$ restent majorées par K, donc $f_h(\theta)$ divise un entier rationnel de l'intervalle fini [1, K] ; on en déduit qu'il existe des indices h et l tels que $f_l(\theta) = e(\theta)f_h(\theta)$, où $e(\theta)$ est une unité, et alors $\varphi(\tilde{L}(e(\theta))) = \varphi(\tilde{L}(f_l(\theta))) - \varphi(\tilde{L}(f_h(\theta))) \neq 0$, comme on voulait.

Le groupe des unités n'est donc fini (et réduit aux racines de 1) que si $r_1 = 1$ et $r_2 = 0$, ce qui donne $n = 1$ et $\theta \in \mathbf{Z}$, ou

bien si Y, $= 0$ et $r_2 = 1$, ce qui donne $n = 2$; dans ce cas, on peut se ramener à $\theta = \sqrt{-D}$, avec D entier rationnel positif non divisible par 4, ou bien à $\theta = (1 + \sqrt{-D})/2$, avec D $\equiv$ 3 (mod 4). On a vu, par exemple, que le groupe des unités de Z[i] est d'ordre 4, tandis que celui de $Z[j]$, j racine cubique de 1, est d'ordre 6 (théorie de Kummer pour A $=$ 3). Lorsque $\theta = \sqrt{D}$, avec D entier rationnel positif sans facteur carré, on a $r_1 = 2$, $r_2 = 0$ et le groupe des unités est de rang $r = 1$; toute unité s'écrit $\pm (T + U\sqrt{D})^n$, où $n \in Z$ et où $T + U\sqrt{D}$ est une *unité fondumentale.* Cela revient à dire que l'équation de Pell $x^2 - Dy^2 = N(x + y\sqrt{D}) = \pm 1$ est résolue en posant $x + y\sqrt{D} = \pm (T + U\sqrt{D})^n$, et on a donc démontré l'existence de solutions pour cette équation. Pour les entiers cyclotomiques, on a, avec les notations antérieures, $r_1 = 0$ et $r_2 = (A\ 1)/2 = \mu$, donc le groupe des unités est de rang $r = \mu - 1$, qui est $\geqslant 1$ à partir de A = 5 ; la difficulté de trouver des lois de réciprocité supérieures sur le modèle des lois pour les degrés 2, 3 et 4 est liée au caractère infini de ces groupes d'unités. Pour A = 5, Y = 1 et on peut prendre le nombre réel a $+ \alpha^4$ comme unité fondamentale ; pour A = 7, $r = 2$ et un système d'unités fondamentales (réelles) est donné par (a $+ \alpha^{-1}$, $\alpha^3 + \alpha^{-3}$). Pour A = 11, $r = 4$, et on a le système fondamental *(a* $+ \alpha^{-1}$, $\alpha^2 + \alpha^{-2}$, $\alpha^4 + \alpha^{-4}$, $\alpha^3 + a^{\ 3}$). Dans le cas cyclotomique général, Kummer considère le sous-groupe du groupe des unités engendré par ± 1, *a* et les unités $(1 - \alpha^j)/(1 - a)$, en prenant $j = $ y', y racine primitive modulo λ et $1 \leqslant i \leqslant \mu = (A - 1)/2$; ce sous-groupe a le même rang $r = \mu - 1$ que le groupe des unités, et le quotient est donc d'ordre fini h_2. En s'inspirant du travail de Dirichlet sur le nombre de classes de formes quadratiques (cf. la partie A ci-dessus - Théorie analytique des nombres), Kummer a donné une formule pour l'ordre h du groupe des classes de diviseurs des entiers cyclotomiques ; il se sert d'une fonction analogue à la fonction zêta (cf. fonction ZÊTA) :

$$\sum_{A} N(A)^{-s} = \prod_{P} (1 - N(P)^{-s})^{-1}$$

(où A parcourt l'ensemble des diviseurs et P celui des diviseurs premiers) et de son comportement pour $s \to 1$, et obtient $h = h_1 h_2$, où le facteur h_2, le plus difficile à calculer, a la signification ci-dessus ; le facteur h_1 est plus explicite :

$$h_1 = \frac{P}{(2\lambda)^{\mu-1}},$$

avec $P = \varphi(\beta)\varphi(\beta^3)\varphi(\beta^5) \dots \varphi(\beta^{\lambda\ 2})$ |, β racine primitive (A $-$ 1)-ième de 1, et φ polynôme défini par $\varphi(x) = 1 + \gamma_1 x + \gamma_2 x^2 + \ + \gamma_{\lambda\ 2} x^{\lambda\ 2}$, où y est une racine primitive modulo A et, pour chaque j, γ_j est le reste de la division de γ^j par λ. Kummer a calculé h_1 pour tous les $\lambda < 100$; la première valeur de A donnant $h_1 \neq 1$ est 23, pour lequel $h_1 = 3$. La croissance de h_1 est très rapide : il vaut 411 322 823 001 pour A = 97 et est équivalent à $\lambda^{\mu/2+1}/2^{\mu-1}\pi^{\mu}$ pour $\lambda \to \infty$; les nombres premiers irréguliers sont ceux pour lesquels h est divisible par λ, et cela équivaut à dire que h_1 est divisible par λ (Kummer a donné un critère simple à vérifier pour cette propriété, au moyen des nombres de Bernoulli).

4. Corps de nombres algébriques

Dedekind (1871, 1893) a étendu les théories précédentes en développant les notions de corps de nombres algébriques et

d'entiers algébriques. Un corps de nombres algébriques est une extension finie du corps Q des nombres rationnels ; un tel corps peut s'écrire K $=$ Q(e), où θ vérifie une équation algébrique irréductible $f(x) = 0$, de degré n, à coefficients rationnels (cf. CORPS), et chacune des n racines complexes de f définit un plongement de K dans le corps C des nombres complexes. On note r_1 le nombre des racines réelles, qui donnent des plongements de K dans R, et $2r_2$ le nombre de racines complexes (paires de racines complexes conjuguées) ; en tant qu'espace vectoriel sur $\mathbf{Q}$, K est de dimension n, avec $1, \theta, \theta^2, \ldots . \theta^{n-1}$ comme base (division euclidienne ; cf. POLYNÔMES). Si g(8) $\in$ K, on appelle conjugués de $g(\theta)$ les n nombres complexes $g(\theta_1), g(\theta_2), ..., g(\theta_n)$, où $\theta_1, \theta_2, \ldots . \theta_n$ sont les racines de $f(x) = 0$; ce sont tous des nombres algébriques.

Entiers algébriques

Parmi les nombres algébriques, les *entiers algébriques* sont définis de manière à former un anneau dont l'intersection avec Q soit réduite à Z ; on veut de plus que tous les conjugués d'un entier algébrique (c'est-à-dire les racines de son équation minimale à coefficients rationnels) soient encore entiers. Alors les coefficients de l'équation minimale d'un entier algébrique sont des entiers algébriques rationnels, c'est-à-dire des éléments de Z ; on définit donc les entiers algébriques comme les racines d'équations à coefficients entiers rationnels, avec un coefficient dominant 1 (cf. ANNEAUX COMMUTATIFS), et il est facile de voir que l'équation minimale d'un tel nombre a encore ses coefficients entiers (donc les entiers algébriques rationnels sont bien les éléments de Z, ce qui généralise le résultat de Théétète cité au début). Pour étudier les entiers du corps K $=$ Q(e), on peut supposer que θ est lui-même entier ; si $\rho = c_0 + c_1\theta + \ldots + c_{n-1}\theta^{n-1}$ est entier, avec $c_i \in \mathbf{Q}$, on a, pour tous ses conjugués $\rho_j = c_0 + c_1\theta_j + \ldots + c_{n-1}\theta_j^{n-1}$. Les coefficients c, sont donc donnés par un système d'équations linéaires de matrice (θ_j^i) $(0 \leqslant i \leqslant n - 1 ; 1 \leqslant j \leqslant n)$; le déterminant de cette matrice est :

$$\prod_{j<k} (\theta_j - \theta_k) = \Delta \neq 0,$$

et on peut résoudre le système par les formules de Cramer qui montrent que l'entier rationnel Δ^2 est un dénominateur commun à tous les c,. Ainsi l'anneau $\mathfrak{o}_K$ des entiers de K est contenu dans le Z-module libre de base $(\theta^i/\Delta^2)_i$ et il est d'indice fini car il contient les θ_i ; par suite, il est lui-même libre et a une base à n éléments $(\omega_1, \omega_2, \ldots . \omega_n)$. La matrice de passage d'une base à une autre appartient à GL (n, Z), et son déterminant vaut $\pm$ 1 ; il en résulte que le déterminant de la matrice $(\omega_j^{(i)})$, où on note $\omega_j^{(1)}, \omega_j^{(2)}, \ldots \omega_j^{(n)}$ les conjugués de ω_j, est défini au signe près par K. Le carré de ce déterminant est un entier rationnel $d \neq 0$, que l'on appelle le *discriminant* de K ; on a :

$$|d| \geqslant \frac{\pi}{3}\left(\frac{3\pi}{4}\right)^{n-1}$$

et il n'y a qu'un nombre fini de corps de discriminant donné (Hermite).

Par exemple, dans le corps Q(i) (avec $i^2 = -1$), les entiers sont de la forme $m + ni$ avec m, n rationnels tels que $m + ni + m - ni = 2m$ et $(m + ni)(m - ni) = m^2 + n^2$ soient entiers ; alors $4m^2 + 4$ n'est un entier, donc aussi $4n^2$, et $2n$ est encore entier. Enfin, la condition que $(2m)^2 + (2n)^2 = 4(m^2 + n^2)$ soit divisible par 4 exige que $2m$ et $2n$ soient pairs, donc m et n sont entiers ; les entiers de Q(i) forment donc l'anneau des entiers de Gauss

$\mathbf{Z}[i]$, de base (1, i). Le discriminant de Q(i) est le carré du déterminant de la matrice :

$$\begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix},$$

soit $(-2i)^2 = $ 4. Pour le corps $\mathbf{Q}(j)$ (avec $j^2 + j + 1 = 0$). les entiers sont $m + nj$, avec $m + nj + m + nj^2 = 2m$ 17 et $(m + nj)(m + nj^2) = m^2 \quad mn + n^2$ entiers, et on voit encore que cela exige m et n entiers ; l'anneau des entiers est $\mathbf{Z}[j]$, de base (1, j), et le discriminant est $(j^2 - j)^2 = $ 3. Passons au cas du corps cyclotomique Q(a), avec $\alpha^{\lambda \ 1} + \alpha^{\lambda-2} + $ + a + 1 = 0, A étant un nombre premier impair (on vient de traiter le cas A = 3) ; si $\rho = c_0 + c_1\alpha + + c_{\lambda-2}\alpha^{\lambda-2}$ est un entier de ce corps, on a :

$$(1-\alpha)\rho = c_0(1-\alpha) + c_1(\alpha - \alpha^2) + \\ .. + c_{\lambda-2}(\alpha^{\lambda-2} - \alpha^{\lambda-1}),$$

et on voit que la somme des conjugués, ou *trace,* de ce nombre vaut $c_0\lambda$, car la trace de α^j est -1 pour $1 \leqslant j \leqslant \lambda - 1$. Comme tous les conjugués $(1 - \alpha^j)\rho_j$ du premier membre sont divisibles par 1 *a,* la trace $c_0\lambda$ l'est aussi ; or c'est un entier rationnel : il est donc divisible par A et c_0 est entier. On peut recommencer ce raisonnement avec $\alpha^{-1}(\rho - c_0)$, qui est encore entier, et montrer que c_1 est entier ; en continuant, on voit que tous les c, sont entiers, et que l'anneau des entiers de Q(a) est $\mathbf{Z}[\alpha]$, avec comme base (1, *a* , ..., $\alpha^{\lambda \ 2}$). Le discriminant est le carré du déterminant **de** la matrice (α^{ij}) $(1 \leqslant i \leqslant$ A $- 1 ; 0 \leqslant j \leqslant \lambda - 2)$, soit :

$$\Big(\prod_{j<i} (\alpha^j - \alpha^i)\Big)$$

$$= (-1)^{(\lambda-1)(\lambda-2)/2} \prod_{j \neq i} (\alpha^j - \alpha^i)$$

$$= (-1)^{(\lambda \quad 1)/2}\, \mathbf{N}(f'(\alpha)),$$

en désignant par f le polynôme :

$$f(x) = x^{\lambda-1} + ... + x + 1 = \frac{X^{"}-1}{X-1};$$

on a $(x-1)f(x) = x^\lambda - 1$, donc $f(x) + (x - 1)f'(x) = \lambda x^{\lambda-1}$, $(\alpha - 1)$ $f'(a) = \lambda\alpha^{\lambda \ -1}$ **et :**

$$\lambda \mathbf{N}(f'(\alpha)) = \mathrm{N}(a - 1)\mathbf{N}(f'(\alpha)) \\ = \lambda^{\lambda-1}(\mathrm{N}\alpha)^{\lambda-1} = \lambda^{\lambda-1}$$

d'où, en définitive $\mathrm{N}(f'(\alpha)) = \lambda^{\lambda-2}$ et $d = (-1)^{(\lambda-1)/2}\lambda^{\lambda-2}$.

A côté de la trace $\mathrm{Tr} f(\theta) = f(\theta_1) + f(\theta_2) \quad + \quad + f(\theta_n)$ d'un élément f (6) de K, on définit sa norme $Nf(\theta) = f(\theta_1)f(\theta_2) f$ *(0,)* ; la trace et la norme sont des nombres rationnels, et ils sont entiers si $f(\theta)$ est entier. Les unités de K sont les entiers de K de norme 1 ; elles forment un groupe multiplicatif pour lequel le théorème de Dirichlet est applicable en toute généralité.

La théorie des idéaux

Dedekind a remplacé la considération des « nombres idéaux », que Kummer n'avait jamais définis comme objets mathématiques, par celle d'objets véritables, qu'il a appelés les *idéaux* du corps K. L'idée est de considérer, au lieu d'un diviseur A et de la congruence $f(\theta) \equiv g(\theta)$ (mod A) qu'il définit dans les entiers algébriques, l'ensemble a de ces entiers qui sont congrus à 0 modulo A, c'est-à-dire l'ensemble des entiers algébriques divisibles par A ; Dedekind a pu caractériser les ensembles **a** d'entiers algébriques ainsi obtenus par les propriétés suivantes : si $f(\theta)$ et g(B) appartiennent à **a,** il en **est** de même de $f(\theta)$ + *g(B)* ; de plus, si $f(\theta) \in \mathbf{a}$ et $h(\theta) \in \mathfrak{o}_K$, alors $f(\theta)h(\theta) \in \mathfrak{a}$. Autrement dit a est un idéal de l'anneau $\mathfrak{o}_K$ des entiers de K (cf. ANNEAUX ET ALGÈBRES); en tant que sous-groupe additif de $\mathfrak{o}_K$, il est

donc libre de rang $\leqslant n$. Tout entier $\rho \in \mathfrak{o}_K$ définit un *idéal principal* (ρ), ensemble des $\rho\sigma$, où σ parcourt $\mathfrak{o}_K$; si a est un idéal quelconque, dire que $\mathfrak{a}$ divise ρ, c'est-à-dire que $\rho \in$ a, revient donc à dire que (p) est contenu dans a. Le plus petit idéal contenant des entiers algébriques $\rho_1, \rho_2, \dots \rho_r$ joue le rôle de leur P.G.C.D., et on le note ($\rho_1, \rho_2, \dots$ p,) ; c'est l'ensemble des combinaisons $\rho_1\sigma_1 + \rho_2\sigma_2 + . + \rho_r\sigma_r$, où $\sigma_1, \sigma_2, \dots \sigma_r$ parcourent $\mathfrak{o}_K$. Tout idéal a peut s'écrire sous la forme ($\alpha_1, \alpha_2, \dots$ a,) en prenant, par exemple, pour $\alpha_1, \alpha_2, \dots$ a, les éléments d'une base de a ; autrement dit, l'anneau $\mathfrak{o}_K$ est nœthérien (cf. ANNEAUXCOMMUTATIFS). Leproduitde deux idéaux a = (a,, $\alpha_2, \dots$ a,) et b = ($\beta_1, \beta_2, \dots \beta_s$) est l'idéal engendré par les produits d'un élément de a et d'un élément de $\mathfrak{b}$, c'est-à-dire ab = $(\alpha_i\beta_j)$ ($1 \leqslant i \leqslant r$, $1 \leqslant j \leqslant s$) ; on dit que l'idéal a divise un idéal i s'il existe un idéal b tel que ab = i, et on écrit alors a i. Un idéal *premier* $\mathfrak{p}$ est un idéal différent de (1) $= \mathfrak{o}_K$ et qui n'est pas divisible par un autre idéal que (1) et $\mathfrak{p}$.

Le fondement de la théorie de Dedekind est le fait que, pour tout idéal $\mathfrak{a}$, il existe un idéal b $\neq$ (0) tel que ab soit principal. On peut construire b en associant, à un système de générateurs α_1, α_2 ,..., a,. de a, le polynôme g(x) = $\alpha_1 x + \alpha_2 x^2 + \alpha_r x^r$ et ses conjugués $g_i(x) = \alpha_1^{(i)}x + + \alpha_r^{(i)}x^r$; le produit F = $g_1 g_2 \dots g_r$ de ces conjugués est un polynôme à coefficients entiers rationnels et il est divisible par g : F = gh, avec $h(x) = \beta_1 x + \beta_2 x^2 + + \beta_s x^s$ polynôme à coefficients entiers algébriques. On pose b = ($\beta_1, \beta_2, \dots \beta^s$) et on montre que ab est l'idéal principal engendré par le P.G.C.D. des coefficients de F en utilisant une extension du lemme de Gauss aux polynômes à coefficients dans $\mathfrak{o}_K$: si un entier algébrique divise tous les coefficients d'un produit de deux polynômes g et h à coefficients dans $\mathfrak{o}_K$, il divise tous les produits d'un coefficient de g par un coefficient de h. On peut alors prouver que l'égalité ab = ai avec a $\neq$ (0) (a, $\mathfrak{b}$, i idéaux) implique b = i ; en effet, en multipliant les deux membres de la première égalité par un idéal convenable on se ramène au cas où a est principal, qui est immédiat. Une autre propriété, qui revient essentiellement à dire que $\mathfrak{o}_K$ est ce que l'on appelle maintenant un *anneau de Dedekind* (cf. ANNEAUX COMMUTATIFS), est la suivante : pour qu'un idéal a divise un idéal i, il faut et il suffit que i soit contenu dans a ; la condition est évidemment nécessaire, et elle est aussi suffisante, car i C a implique ib C ab pour tout idéal b, ce qui permet de se ramener au cas facile où a est principal. Ainsi l'idéal a + b engendré par deux idéaux a et b est aussi le plus grand idéal qui divise à la fois a et b (supposés non tous les deux nuls) ; en utilisant ce P.G.C.D. comme dans l'arithmétique élémentaire, on montre que si un idéal premier $\mathfrak{p}$ divise un produit d'idéaux. il divise l'un des facteurs, et on en déduit que tout idéal non nul et distinct de (1) s'écrit, d'une manière unique, comme produit d'idéaux premiers.

Pour tout idéal non nul a, l'anneau quotient $\mathfrak{o}_K/\mathfrak{a}$ est fini ; en effet, si α est un élément non nul de a, (N(a)) C (α) C $\mathfrak{a}$, donc $\mathfrak{o}_K/\mathfrak{a}$ est un quotient de $\mathfrak{o}_K/(N(\alpha))$, qui est visiblement fini. La norme de a est. par définition, le nombre Na d'éléments de $\mathfrak{o}_K/\mathfrak{a}$; lorsque a = (α) est principal. N est la valeur absolue de N(a). Le théorème chinois signifie que la norme est multiplicative : N(ab) = $N\mathfrak{a} \cdot$ Nb. Si $\mathfrak{p}$ est un idéal premier, il divise au moins un entier rationnel (la norme d'un de ses éléments). donc, si $\mathfrak{p} \neq$ (0), il divise un nombre premier rationnel p ; alors $N\mathfrak{p}$ divise

$N\mathfrak{p} = p^n$, et on a donc $N\mathfrak{p} = p^f$, où f est un entier $\leqslant n$, que l'on appelle le *degré* de $\mathfrak{p}$. L'anneau fini $\mathfrak{o}_K/\mathfrak{p}$ est intègre, donc c'est un corps à p^f éléments. Pour tout $\alpha \in \mathfrak{o}_K$, $\alpha^{N\mathfrak{p}} \equiv \alpha$ (mod $\mathfrak{p}$) et. si $\alpha^p \equiv \alpha$ (mod $\mathfrak{p}$), α est congru modulo $\mathfrak{p}$ à un entier rationnel. Dans le cas où le corps $K = \mathbf{Q}(\theta)$ est galoisien sur $\mathbf{Q}$, c'est-a-dire que le polynôme minimal $f(x)$ de θ se décompose sur K en facteurs du premier degré, les n plongements de K dans C ont la même image, que l'on peut identifier à K, et un idéal $\mathfrak{a}$ de K donne n images $\mathfrak{a}_1, \mathfrak{a}_2, \ldots, \mathfrak{a}_n$ qui sont les conjugués de $\mathfrak{a}$; le produit $\mathfrak{a}_1 \mathfrak{a}_2 \ldots \mathfrak{a}_n$ est l'idéal principal de K engendré par $N\mathfrak{a}$. En particulier, la norme p^f d'un idéal premier $\mathfrak{p}$ est le produit des idéaux conjugués de $\mathfrak{p}$; les idéaux premiers divisant p sont donc conjugués de $\mathfrak{p}$, et si aucun d'eux ne divise p avec une multiplicité $\geqslant 2$, f divise n et il y a n/f conjugués distincts de $\mathfrak{p}$, chacun répété f fois.

Considérons par exemple un corps *quadratique* $\mathbf{Q}(\sqrt{D})$, où D est un entier rationnel sans facteur carré ; dans ce cas, $\theta = \sqrt{D}$ est racine de $f(x) = x^2 - D = 0$, dont les racines complexes sont $\pm \sqrt{D}$. Le corps est galoisien, et les conjugués de $x + y\sqrt{D}$ sont les nombres $x \pm y\sqrt{D}$; la trace et la norme sont respectivement $2x$ et $x^2 - y^2D$, et les entiers du corps sont caractérisés par le fait que ces deux nombres rationnels sont entiers. En raisonnant comme plus haut, on voit que cela signifie que x et y sont entiers si $D \equiv 2$ ou 3 (mod 4) ; mais, si $D \equiv 1$ (mod 4). cela signifie que $x = u/2$ et $y = v/2$, où u et v sont des entiers de même parité. Dans le premier cas, une base des entiers est $(1, \sqrt{D})$, et le discriminant vaut 4 D, tandis que dans le second cas une base des entiers est $(1, (1 + \sqrt{D})/2)$ et le discriminant est D. Soit p un nombre premier rationnel qui ne divise pas le discriminant d ; si p se décompose dans K, on a $(p) = \mathfrak{p}\mathfrak{p}'$ où $\mathfrak{p}$ et $\mathfrak{p}'$ sont des idéaux premiers de degré 1 conjugués l'un de l'autre, et tout entier de K est congru modulo $\mathfrak{p}$ à un entier rationnel ; on en déduit que D est congru modulo p (resp. modulo $4p$) à un carré, c'est-à-dire que le discriminant d est un carré modulo $4p$, donc aussi modulo $4p$. Inversement, si $d \equiv x^2$ (mod $4p$), le nombre $(x + \sqrt{d})/2$ est un entier de K (sa trace est x et sa norme un multiple entier de p) qui n'est pas divisible par p (sinon son conjugué $(x - \sqrt{d})/2$ le serait aussi, donc aussi $\sqrt{d}$, et d serait divisible par p^2) ; comme p divise :

$$\frac{x + \sqrt{d}}{2} \, \frac{x - \sqrt{d}}{2}$$

sans diviser aucun facteur, il n'est pas premier dans $\mathfrak{o}_K$, et il se décompose. Si maintenant q est un facteur premier impair de d, on vérifie que (q) est le carré de l'idéal premier $\mathfrak{q} = (q, (d + \sqrt{d})/2)$, qui est égal à son conjugué. Si enfin d est pair, on a $2 = (2, \sqrt{D})^2$ si $D \equiv 2$ (mod 4) et $2 = (2, 1 + \sqrt{D})^2$ si $D \equiv 3$ (mod 4). Comme d est toujours congru à 0 ou à 1 modulo 4, son caractère quadratique modulo $4p$ est le même que modulo p si $p \neq 2$; les nombres premiers rationnels se rangent donc en trois catégories :

1. Ceux qui ne divisent pas d et modulo lesquels d est un carré, qui se décomposent en produit de deux idéaux premiers distincts conjugués.

2. Ceux qui ne divisent pas d et modulo lesquels d n'est pas un carré, qui restent premiers dans K.

3. Ceux qui divisent d, qui sont des carrés d'idéaux premiers égaux à leur conjugué. (Pour $p = 2$, on teste le caractère de d mod 8.)

Cela généralise ce qu'on avait observé pour Q(i), Dedekind a démontré un résul-

tat plus général, qui englobe aussi les résultats de Kummer sur la décomposition des nombres premiers rationnels dans les corps cyclotomiques : si K = $\mathbf{Q}(\theta)$ est un corps de nombres algébriques engendré par un entier algébrique θ d'équation minimale $f(\theta) = 0$ et si p est un nombre premier qui ne divise pas l'indice $C(\theta) = (\mathfrak{o}_K : Z[\theta])$, à une décomposition modulo p :

$$f(x) \equiv \prod_{i=1} (P_i(x))^{e_i} \pmod p$$

de $f(x)$ en produit de polynômes irréductibles P, (distincts modulo p) correspond une décomposition :

$$(p) = \prod_{i=1}^{r} \mathfrak{p}_i^{e_i}$$

en produit d'idéaux premiers $\mathfrak{p}_i$ distincts, les multiplicités e_i étant les mêmes. Pour chaque i, on dit que e_i est l'indice de ramification de $\mathfrak{p}_i$ et, en considérant la norme $N\mathfrak{p} = p^n$ de p, on voit que :

$$n = \sum_{i=1} e_i f_i,$$

où les f_i sont les degrés des $\mathfrak{p}_i$. Malheureusement, il y a des cas où on ne peut pas obtenir la décomposition de p par le résultat précédent (lorsque p divise $C(\theta)$ quel que soit le choix de θ). Les idéaux premiers ramifiés $\mathfrak{p}$, c'est-à-dire ceux qui divisent un nombre premier rationnel p avec un indice de ramification $e \geqslant 2$, sont les diviseurs d'un idéal $\mathfrak{d}_K$ bien déterminé, dont la norme est d |; on appelle $\mathfrak{d}_K$ la *différente* de K.

idéaux fractionnaires ; classes d'idéaux

On appelle idéal fractionnaire de K un sous-o,-module non nul a de K tel qu'il existe un entier non nul δ vérifiant $\delta\mathfrak{a} \subset \mathfrak{o}_K$; alors a est engendré, comme $\mathfrak{o}_K$-module, par un nombre fini d'éléments de K. Les idéaux fractionnaires forment un groupe pour la multiplication, avec $(1) = \mathfrak{o}_K$ comme élément unité ; c'est un groupe commutatif libre avec comme base l'ensemble des idéaux premiers.

Dedekind définit l'équivalence des idéaux a et b par l'existence d'un même idéal i $\neq (0)$ tel que ai et bi soient tous deux principaux ; il revient au même de dire que l'idéal fractionnaire quotient $\mathfrak{a}\mathfrak{b}^{-1}$ est principal. Autrement dit, le groupe des classes d'idéaux est le groupe quotient du groupe des idéaux fractionnaires par le sous-groupe des idéaux fractionnaires principaux (c'est-à-dire engendrés par un seul élément). Par une méthode analogue à celle de Kummer, Dedekind montre, dans le cas général, que le groupe des classes d'idéaux est fini ; en utilisant la fonction :

$$\zeta_K(s) = \sum_{\mathfrak{a} \neq 0} (N\mathfrak{a})^{-s}$$

et son comportement pour $s \rightarrow 1$, il établit un lien remarquable entre le nombre h des classes d'idéaux et la « densité » des idéaux.

Par ailleurs, Dedekind a montré que la classification des formes quadratiques binaires développée par Gauss était essentiellement équivalente à celle des idéaux du corps quadratique de même discriminant ; la loi de groupe définie par Gauss au moyen de la composition des formes sur l'ensemble des classes de formes quadratiques de discriminant donné correspond à celle du groupe des classes d'idéaux.

5. Corps de classes

La difficile théorie du corps de classes tire son origine de plusieurs résultats établis au

cours du XIX' siècle. Nous avons vu que Gauss avait associé, à tout nombre premier impair p, une somme :

$$\sum \left(\frac{m}{p}\right) r^m \in \mathbf{Q}(r),$$

corps des racines p-ièmes de 1, dont le carré est $(-1)^{(p-1)/2}p$; le sous-corps de Q(Y) engendré par la somme de Gauss est donc isomorphe au corps quadratique $Q(\sqrt{(-1)^{(p-1)/2}p})$. Kronecker a obtenu une vaste généralisation de ce résultat (la démonstration complète est due à Weber) : tout corps K de nombres algébriques dont le groupe de Galois sur Q est commutatif se plonge dans un corps cyclotomique. La théorie de la multiplication complexe des fonctions elliptiques a ensuite conduit Kronecker à formuler une conjecture analogue pour les corps de nombres algébriques K contenant un corps quadratique imaginaire k et tel que le groupe de Galois $G(K/k)$ soit commutatif (« rêve de jeunesse de Kronecker », 1857) ; dans cette conjecture, qui n'a été complètement démontrée qu'en 1920, les fonctions elliptiques admettant de la multiplication complexe par certains entiers de k jouent le rôle que jouait I'exponentielle imaginaire pour les racines de 1.

En étendant la théorie de Kummer aux corps cyclotomiques Q(o) = K, où α est une racine m-ième de 1, m entier quelconque. on constate que la décomposition d'un nombre premier rationnel p qui ne divise pas md (d discriminant de K) ne dépend que de l'ordre f de p modulo m : p se décompose en $e = \varphi(m)/f$ facteurs premiers idéaux distincts de degré f dans K, où $\varphi(m)$ est l'indicateur d'Euler (cf. DIVISIBILITÉ). En particulier, si $p \equiv 1 \pmod m$, c'est le produit de $\varphi(m)$ idéaux premiers de degré 1 ; au moyen de la fonction ζ_K, on peut montrer que l'ensemble des idéaux premiers de degré 1 de K est infini, et il en résulte qu'il y a une infinité de nombres premiers dans la progression arithmétique de raison m qui contient 1 (cf. la partie A ci-dessus - Théorie analytique des nombres). Weber a essayé de généraliser ce genre de considérations en remplaçant Q par un corps de nombres algébriques k et m par un idéal $\mathfrak{m}$; il considère le groupe $A_{\mathfrak{m}}$ des idéaux fractionnaircs dc k premiers à $\mathfrak{m}$ et un sous-groupe $H_{\mathfrak{m}}$ d'indice fini h' formé d'idéaux principaux dans $A_{\mathfrak{m}}$. Il fait alors les hypothèses suivantes :

a) Les idéaux entiers de k sont « également distribués » dans les classes de $A_{\mathfrak{m}}/H_{\mathfrak{m}}$ (comme ils le sont dans les classes d'idéaux habituelles).

b) Il existe une extension K de k de degré $\leqslant h'$ telle que lcs idéaux premiers de $H_{\mathfrak{m}}$ de degré 1 se décomposent complètement dans K (c'est-à-dire en produit d'idéaux distincts tous de degré 1); l'extension K s'appelle un corps de classes pour k.

Weber montre alors que chaque classe de $A_{\mathfrak{m}}/H_{\mathfrak{m}}$ contient une infinité d'idéaux du premier degré. Dans le cas où $k = \mathbf{Q}$, on peut prendre pour $H_{\mathfrak{m}}$ le groupe des idéaux engendré par un nombre congru à 1 modulo MI ; dans ce cas, Weber a aussi établi quc lc groupe de Galois de $\mathbf{Q}(\zeta_m) = K$ sur Q s'identifie à $A_{\mathfrak{m}}/H_{\mathfrak{m}}$ (ζ_m racine m-ième de 1). et qu'a chaque sous-groupe $H'_{\mathfrak{m}} \supset H_{\mathfrak{m}}$ correspond un corps de classes $K' \subset \mathbf{Q}(\zeta_m)$ de groupe de Galois $A_{\mathfrak{m}}/H_{\mathfrak{m}}$. Dans le cas général, en supposant l'existence du corps de classes K, Weber a seulement démontré que son degré sur k est égal a h' et qu'il est galoisien sur k. Revenant au cas particulier $k = Q$. si L est une extension abélienne quelconque de $\mathbf{Q}$, ellc sc plonge dans un corps $\mathbf{Q}(\zeta_m)$ (théorème de Kronecker-Weber) et

correspond donc à un groupe d'idéaux principaux H'_m, tel que $H_m \subset H'_m \subset A_m$, et que Gal(L/Q) $\simeq A_m/H'_m$; l'entier m n'est pas unique, mais il admet une valeur minimale dont toutes les autres sont des multiples (le conducteur de L). Weber a encore formulé des conjectures qui étendent ces énoncés au cas où Q est remplacé par un corps de nombres algébriques ***k***.

Hilbert a abordé la théorie du corps de classes d'un autre point de vue, à partir de la théorie des formes quadratiques, et il est parvenu à construire certains corps de classes, correspondant au groupe de classes d'idéaux A/H_+, où A est le groupe des idéaux fractionnaires d'un corps de nombres algébriques k et H_+ le groupe des idéaux principaux engendré par des entiers de k qui sont positifs dans tous les plongements réels de ***k***. Le corps de classes de Hilbert est unique, et son groupe de Galois sur ***k*** est A/H_+ ; dans ce corps, tous les idéaux de $\mathfrak{o}_k$ deviennent principaux.

Frobenius (1896) a introduit un objet important dans la théorie du corps de classes : l'automorphisme de Frobenius, ainsi construit. On considère une extension galoisienne finie K de Q et un idéal premier $\mathfrak{p}$ de K, divisant un nombre premier rationnel p et non ramifié ; alors le sous-groupe $Z(\mathfrak{p})$ du groupe de Galois Gal $(\mathbf{K}/\mathbf{Q})$ formé des automorphismes de K qui laissent $\mathfrak{p}$ invariant (« groupe de décomposition ») s'identifie au groupe de Galois du corps résiduel $\mathfrak{o}_K/\mathfrak{p}$ sur $\mathbf{Z}/(p)$. À l'élévation à la puissance p-ième, qui est un automorphisme de $\mathfrak{o}_K/\mathfrak{p}$, correspond donc un élément $((\mathbf{K}/\mathbf{Q})/\mathfrak{p})$ de $Z(\mathfrak{p})$, qui est par définition l'automorphisme de Frobenius ; pour $\sigma \in$ Gal $(\mathbf{K}/\mathbf{Q})$ quelconque, $((\mathbf{K}/\mathbf{Q})/(\sigma\mathfrak{p}))$ est conjugué de $((\mathbf{K}/\mathbf{Q})/\mathfrak{p})$ sous l'action de σ. Comme les idéaux premiers de K qui divisent p sont tous conjugués, les automorphismes de Frobenius qui leur correspondent appartiennent à une même classe de conjugaison dans Gal $(\mathbf{K}/\mathbf{Q})$, et en particulier ils sont égaux si ce dernier groupe est commutatif. Čebotarëv (1925) a pu montrer que, pour toute classe de conjugaison à m éléments dans Gal $(\mathbf{K}/\mathbf{Q})$, la densité des nombres premiers p qui donnent des automorphismes de Frobenius appartenant à cette classe est m/n, où $n = (K : Q)$ est le degré de K. La loi de réciprocité d'Artin (1927 ; la formulation d'Artin vaut en fait pour un corps de base ***k*** général, et pas seulement pour Q) signifie que, pour une extension abélienne L de $\mathbf{Q}$, l'automorphisme de Frobenius $((\mathrm{L}/\mathbf{Q})/p)$ correspondant à un nombre premier p est l'identité exactement dans le cas où p appartient au sous-groupe H'_m correspondant (où m est le conducteur de L).

6. Idèles et adèles

Dans ses recherches sur les formes quadratiques à coefficients dans un corps de nombres algébriques ***k***, en vue d'étendre un résultat de Minkowski, Hilbert avait été conduit à considérer simultanément des congruences modulo les puissances des idéaux premiers du corps, et les équations correspondantes dans R ou dans C, provenant des divers plongements de ***k***; il appelait *place* de ***k*** un idéal premier de k, ou bien un plongement de ***k*** dans R ou dans C, ces dernières places étant qualifiées de « places à l'infini ». Takagi (1920), dans ses démonstrations des conjectures de Weber en théorie du corps de classes, a modifié la notion de diviseur telle que nous l'avons introduite plus haut, de manière à inclure les places à l'infini : selon Takagi, un diviseur est un symbole $\mathfrak{m} = \mathfrak{p}_1^{n_1} \mathfrak{p}_2^{n_2} \dots \mathfrak{p}_r^{n_r}$, avec $n_i \in \mathbf{N}$, les $\mathfrak{p}_i$

étant des places finies (= idéaux premiers) ou non. Dans les diviseurs fractionnaires, on admet des exposants n_i négatifs, et on a ainsi un groupe multiplicatif; à un diviseur entier m, on associe le groupe A,,, des diviseurs fractionnaires premiers à m, et le sous-groupe Hm des diviseurs principaux congrus à 1 modulo m, c'est-à-dire congrus à 1 modulo $\mathfrak{p}_i^{n_i}$ pour tout i tel que $\mathfrak{p}_i$ soit un idéal premier, et d'image positive pour toute place à l'infini réelle p,. Les groupes de classes d'idéaux de Weber sont alors remplacés par les groupes de classes de diviseurs $A_\mathfrak{m}/H_\mathfrak{m}$ et leurs quotients de la forme $A_\mathfrak{m}/H_\mathfrak{m} . N_{K/k}(A_\mathfrak{m}(K))$, où K est une extension galoisienne finie de k, A,,,(K) est le groupe des diviseurs fractionnaires de K premiers à $\mathfrak{m}$, et $N_{K/k}$ est la « norme relative » ; si l'ordre du groupe quotient précédent est égal au degré (K : k), K est un corps de classes au sens de Takagi, et son groupe de Galois sur k est isomorphe à ce quotient. Toute extension abélienne de k est un corps de classes pour un certain diviseur m, que l'on peut choisir minimal (le « conducteur » de K).

Une théorie analogue, mais beaucoup plus simple, a été développée par Hasse (1929-1930), en considérant, au lieu du corps de nombres algébriques k son complété $k_\mathfrak{p}$ pour la valuation associée a l'idéal premier p ; ce corps $k_\mathfrak{p}$ est une extension finie du corps p-adique $\mathbf{Q}_p$, où p est le nombre premier que p divise (cf. la partie B ci-dessus - Nombresp-adiques), et il a des propriétés analogues : son anneau $\mathfrak{o}_\mathfrak{p}$ des entiers, éléments de valuation $\geqslant 0$, est principal ct il a un seul idéal premier non nul, engendré par p. La théorie du *corps de classes local* de Hasse établit une correspondance bijective entre les sous-groupes d'indice fini H du groupe multiplicatif de $k_\mathfrak{p}$ et les extensions abéliennes finies $K_\mathfrak{p}$ de $k_\mathfrak{p}$ (ce sont aussi des corps locaux, extensions finies de $\mathbf{Q}_p$, et on note $\mathfrak{P}$ l'unique idéal premier non nul) ; pour une telle extension, H est l'image du groupe multiplicatif de $K_\mathfrak{p}$ par la norme relative $N_{K_\mathfrak{p}/k_\mathfrak{p}}$. Les démonstrations de Hasse étaient fondées sur la théorie « globale » de Takagi, mais Chevalley (1933) est parvenu à un exposé autonome de la théorie locale. Il eut ensuite l'idée de récupérer la théorie globale a partir de la théorie locale (1936-1940), en remplaçant les diviseurs de Takagi par les *idèles*. Un idèle de k est un élément $(\xi_\mathfrak{p})_\mathfrak{p}$ du produit des groupes multiplicatifs de tous les complétés $k_\mathfrak{p}$ de k, $\mathfrak{p}$ variant dans l'ensemble de toutes les places, y compris les places à l'infini ; pour ces dernières, le complété est R si la place est réelle, et C si la place est imaginaire (d'après un théorème d'Ostrowski (1935), toutes les valeurs absolues possibles sur k sont équivalentes à l'une de celles qui sont définies par les places ; cf. algèbre TOPOLOGIQUE). On impose, de plus, que $v_\mathfrak{p}(\xi_\mathfrak{p}) = 0$ sauf pour un nombre fini de places finies p, en notant $v_\mathfrak{p}$ la valuation correspondante ; l'ensemble *I(k)* des idèles de k est un sous-groupe du produit :

$$\prod_\mathfrak{p} k_\mathfrak{p}^\times$$

des groupes multiplicatifs. Le groupe multiplicatif k^* de k se plonge dans *I(k)*, chaque $\xi \in k$ donnant pour image l'idèle $(\xi_\mathfrak{p})$ tel que $\xi_\mathfrak{p} = \xi$ pour toute place p ; on a un homomorphisme de I(k) sur le groupe des idéaux fractionnaires de k, qui transforme l'idèle $(\xi_\mathfrak{p})$ en l'idéal :

$$\prod_\mathfrak{p} \mathfrak{p}^{v_\mathfrak{p}(\xi_\mathfrak{p})}$$

(produit étendu aux places finies). Le noyau de cet homomorphisme est l'ensemble U(k) des idèles $(\xi_\mathfrak{p})$ tels que $\xi_\mathfrak{p}$ soit

une unité de $k_{\mathfrak{p}}$ (élément de valuation 0) pour toute place finie $\mathfrak{p}$; il en résulte que le groupe des classes d'idéaux est isomorphe au quotient $I(k)/k^{*}U(k)$. Lorsque K est une extension abélienne finie de k, Chevalley définit un homomorphisme de norme $N_{K/k} : I(K) \rightarrow I(k)$ en combinant les normes relatives locales ; à tout idèle a, il associe un symbole $(a, K/k) \in \mathrm{Gal}(K/k)$, égal au symbole d'Artin :

$$\left(\frac{K/k}{\mathfrak{a}}\right),$$

où $\mathfrak{a}$ est l'idéal associé à un idèle $a' \in k^{*}a$ et congru à 1 modulo le conducteur de K. La loi de réciprocité d'Artin signifie que les idèles a tels que $(a, K/k)$ soit l'identité sont les éléments de $k^{*} \cdot N_{K/k}(I(K))$, de sorte que $\mathrm{Gal}(K/k)$ est isomorphe à $I(k)/k^{*} \cdot N_{K/k}(I(K))$.

La théorie multiplicative des idèles doit être complétée par une théorie additive, celle des *adèles*, introduits par A. Weil : un adèle de k est un élément $(\xi_{\mathfrak{p}})_{\mathfrak{p}}$ du produit :

$$\prod_{\mathfrak{p}} k_{\mathfrak{p}}$$

tel que $v_{\mathfrak{p}}(\xi_{\mathfrak{p}}) \geqslant 0$ sauf pour un nombre fini de places finies $\mathfrak{p}$. L'ensemble A(k) des adèles est un sous-anneau du produit :

$$\prod_{\mathfrak{p}} k_{\mathfrak{p}},$$

dont I(k) est le groupe multiplicatif des éléments inversibles ; on plonge le corps k dans A(k) comme $k^{\times}$ dans I(k). En plus de ces structures algébriques, A(k) a une topologie localement compacte, compatible avec sa structure d'anneau, et qui provient du fait que chaque $\mathfrak{o}_{\mathfrak{p}}$ est un anneau *compact*. donc le produit :

$$\prod_{\mathfrak{p}} \mathfrak{o}_{\mathfrak{p}},$$

est aussi compact (théorème de Tychonoff ; cf. TOPOLOGIE GÉNÉRALE) ; pour obtenir A(k), on remplace un nombre fini de facteurs $\mathfrak{o}_{\mathfrak{p}}$ par le corps localement compact $k_{\mathfrak{p}}$ et on ajoute un nombre fini de facteurs égaux à R ou à C. La topologie induite sur I(k) n'est pas compatible avec la structure de groupe, et on doit la remplacer par la topologie du graphe de l'application $a \mapsto a^{-1}$ ($a \in I(k)$), considéré comme sous-espace de A(k) X A(k). Les résultats essentiels concernant ces topologies sont les suivants : l'image de k (resp. k^{*}) dans A(k) (resp. I(k)) est *discrète et* le quotient $A(k)/k$ (resp. $I(k)/k^{*}$) est *compact* ; de ces résultats, on peut déduire sans peine la finitude du nombre de classes d'idéaux et le théorème des unités de Dirichlet. D'une manière plus générale, on peut associer, **à tout** groupe algébrique linéaire G (cf. GÉOMÉTRIE ALGÉBRIQUE) défini sur le corps k, un groupe localement compact G_{A}, adélisé de G, construit comme I(k) à partir du groupe multiplicatif ; Tamagawa et A. Weil ont montré que, dans le cas d'un groupe semi-simple, le groupe G_{k} des points rationnels sur k de G se plonge dans G_{A} comme sous-groupe discret et que l'espace homogène quotient G_{A}/G_{k} est de volume fini pour une mesure semi-invariante. Le calcul de ce volume, par exemple pour le cas où G est le groupe orthogonal associé à une forme quadratique, est équivalent aux résultats de Siegel sur le nombre de représentations d'une forme quadratique à coefficients entiers par un genre donné de formes (cf. formes QUADRATIQUES).

CHRISTIAN HOUZEL

Bibliographie

Y. AMICE, *Les Nombres p-adiques*, Presses universitaires de France, Paris, 1975 / E. ARTIN & J. TATE.

Class Field Theory, Benjamin, New York, 1968, rééd., Addison-Wesley, Redding (Mass.), 1990 / Z. 1. BOREVIČ & I. R. ŠAFAREVIČ, *Number Theory*, Acad. Press, New York, 1966, reprod. en fac-sim., J. Gabay, Sceaux. 1993 / J. W. S. CASSELS & M. J. TAYLOR, *Algebraic Number Theory*, Cambridge Univ. Press, New York, 1992 / J. DIEUDONNÉ et al., *Abrégé d'histoire des mathématiques*, Hermann. Paris, nouv. éd., 1986 / H. M. EDWARDS, *Fermat's Last Theorem*, Springer, New York, 1989 / D. HILBERT, *Théorie des corps de nombres algébriques*, J. Gabay, 1991 / K. IRELAND & M. ROSEN, *A Classical Introduction to Modern Introduction to Modern Number Theory*, *ibid.*, 1990 / S. LANG. *Algebraic Number Theory*, rééd., Springer, 1986 / P. RIBENBOÏM, *The Book of Prime Number Records*, Springer, 1989 / P. SAMUEL, *Théorie algébrique des nombre.~*, Hermann, 1967 / J.-P. SERRE, *Corps locaux*, *ibid.*, 3' éd. 1980 / A. WEIL, *Basic Number Theory*, Springer, Berlin, 1985.

NOMBRES COMPLEXES

→ COMPLEXES NOMBRES

NOMBRES TRANSCENDANTS

→ TRANSCENDANTS NOMBRES

NORMÉES ALGÈBRES

Au point dc rcncontrc dc deux types de structures, structures algébriques et structures topologiques, les algèbres normées jouent un rôle important dans de nombrcux domaincs dc l'analyse mathématique. Développée à partir de 1940 environ, essentiellement par des mathématiciens soviétiques (I. M. Gelfand, M. A. Naimark, D. A. Raikov, G. E. Šylov), la théorie des algèbres normées avait primitivement pour objet de placer dans un cadre abstrait et général l'étude de certaines algèbres normées particulières (en l'occurrence, les algèbres de convolution de fonctions intégrables pour une mesure de Haar d'un groupe localement compact) en isolant leurs propriétés les plus marquantes et les plus caractéristiques.

On reconnaît là le processus d'axiomatisation, qui a été si souvent utilisé en mathématiques et qui est si riche de conséquences.

Historiquement issue de l'analyse harmonique (dont l'un des principaux objets est précisément l'étude de l'algèbre de convolution des fonctions intégrables sur un groupe), la théorie des algèbres normées permit par la suite d'obtenir de nouveaux résultats aussi bien en analvse harmonique que dans d'autres branches de l'analyse (cf. analyse HARMONIQUE).



1. La notion d'algèbre normée

Définition

Une algèbre normée est un ensemble muni à la fois d'une structure d'espace vectoriel sur le corps des nombres complexes, d'une structure d'anneau et d'une norme (cf. espaces vectoriels NORMÉS ; ANNEAUX ET ALGÈBRES).

Plus précisément, notons C le corps des nombres complexes. Un ensemble A est alors une algèbre normée si les conditions suivantes sont réunies :

a) On définit sur A deux lois de composition interne, addition et multiplication, qui munissent A d'une structure d'anneau ;

b) On définit une loi de composition externe, multiplication par les scalaires complexes, qui, jointe à la loi interne d'addition, munit A d'une structure d'espace vectoriel sur C ;

c) Les structures d'anneau et d'espace vectoriel sont compatibles en ce sens que, quels que soient les éléments A de C et les éléments a et *b* de A, on a :

$$\lambda(ab) = (\lambda a)b = a(\lambda b);$$

d) On définit sur A une norme, c'est-à-dire une application $x \longmapsto \| x \|$ de A dans l'ensemble des nombres réels positifs telle que, quels que soient les éléments λ de C et les éléments *a, b* et c de A, on ait :

$$\|a\| = 0,$$

si et seulement si a = 0, élément neutre de l'addition dans A,

$$\|a + b\| \leqslant \|a\| + \|b\|,$$
$$\|\lambda a\| = |\lambda| \, \|a\|,$$
$$\|ab\| \leqslant \|a\| \, \|b\|;$$

e) La distance déduite de la norme (la distance de deux éléments *a* et *b* étant, par définition, $\|a - b\|$) munit A d'une structure d'espace complet (cf. espaces MÉTRIQUES, chap. 3).

Pour cette raison, les algèbres normées sont fréquemment appelées *algèbres de Banach*, par analogie avec les espaces vectoriels normés complets, dits espaces de Banach.

Si la multiplication interne est commutative, on parle d'algèbre normée commutative. Si la multiplication interne possède une unité, on parle d'algèbre normée unitaire.

Exemples

Indiquons trois types fondamentaux d'algèbres normées.

(1) Soit X un espace topologique, et soit A l'ensemble des fonctions continues et bornées sur X, muni des opérations usuelles et de la norme :

$$\|a\| = \sup_{x \in X} |a(x)|;$$

c'est une algèbre normée commutative unitaire.

(2) Soit E un espace de Banach et soit A = L(E) l'ensemble des applications linéaires continues de E dans lui-même. L'addition et la multiplication par les scalaires sont définies de manière usuelle ; la multiplication interne est la composition des opérateurs linéaires. Quant à la norme, elle est définie par :

$$\|a\| = \sup_{\substack{x \in E \\ \|x\|_E = 1}} \|a(x)\|_E;$$

c'est la norme habituelle des opérateurs. A est ainsi muni d'une structure d'algèbre normée unitaire (l'unité de A est l'opérateur I_E, identité de E dans E), non commutative si E est de dimension supérieure à 1.

(3) G est un groupe localement compact et μ est une mesure de Haar à gauche sur G (cf. analyse HARMONIQUE, chap. 4). Rappelons que c'est une mesure telle que l'on ait, pour toute fonction intégrable *f* et pour tout élément *t* de G,

$$\int f \, d\mu = \int {}_t f \, d\mu,$$

la fonction ${}_t f$, translatée de *f* à gauche par *t* étant définie par :

$${}_t f(x) = f(t^{-1}x).$$

A est l'espace de Banach $L^1(\mu)$ des fonctions μ-intégrables, muni de sa norme :

$$\|f\| = \int |f| \, d\mu;$$

la multiplication interne est l'opération de convolution, notée « $*$ », définie par :

$$(f*g)(x) = \int f(y)\,g(y^{-1}x)\,d\mu(y),$$

formule ayant un sens « μ-presque-partout ».

On a ainsi défini une algèbre normée, commutative lorsque le groupe G est commutatif, unitaire lorsque le groupe G est muni de la topologie discrète, Cet exemple, auquel il a été fait allusion dans l'introduction, est à l'origine de toute la théorie.

2. Les algèbres normées commutatives

Nous allons examiner quelques propriétés fondamentales des algèbres normées en présentant d'abord la théorie dans le cas des algèbres normées commutatives et unitaires.

Idéaux maximaux et caractères

L'étude des idéaux maximaux est sans doute l'outil le plus puissant pour obtenir des propriétés des algèbres normées commutatives unitaires,

Indiquons brièvement qu'un *idéal* d'une algèbre normée commutative A est une partie 1 de A qui est un sous-espace vectoriel de A et qui, d'autre part, contient l'élément ab dès que a est un élément de I et b un élément quelconque de A. Évidemment A est un idéal (peu intéressant !) de 4. Un idéal est dit *maximal* s'il n'est contenu strictement dans aucun idéal autre que l'algèbre A elle-même.

On appelle *caractère* de l'algèbre normée commutative unitaire A tout homomorphisme non identiquement nul de A dans C : autrement dit, un caractère de A est une fonction χ définie sur A, à valeurs complexes, non identiquement nulle, telle que, quels que soient λ dans C, et a et b dans A, on ait :

$$\chi(\lambda a) = \lambda\chi(a),$$
$$\chi(a+b) = \chi(a) + \chi(b), \qquad \chi(ab) = \chi(a)\chi(b).$$

Il est facile de vérifier que le noyau d'un caractère, c'est-à-dire l'ensemble des éléments de A où s'annule ce caractère, est un idéal maximal. En fait, caractères et idéaux maximaux satisfont aux propriétés suivantes :

a) Tout idéal propre (c'est-à-dire distinct de A) est contenu dans au moins un idéal maximal ;

b) Tout idéal maximal est fermé pour la topologie définie par la norme sur A ;

c) Tout idéal maximal est le noyau d'un caractère bien déterminé, et tout caractère admet pour noyau un idéal maximal : cela établit une correspondance biunivoque entre les idéaux maximaux et les caractères.

Les deux dernières propriétés entraînent le fait remarquable que, pour une algèbre normée commutative unitaire, tout caractère (défini uniquement par des propriétés algébriques) est automatiquement continu.

Spectre et transformation de Gelfand

L'ensemble des caractères de A est appelé *spectre* de A : nous noterons A(A) cet ensemble.

À tout élément a de A on peut associer une fonction â, appelée *transformée de Gelfand* de a, définie sur A(A), à valeurs complexes : la valeur de â au point χ de Δ(A) est simplement la valeur prise par le caractère χ au point a de A :

$$\hat{a}(\chi) = \chi(a).$$

Il existe sur Δ(A) une topologie d'espace compact et une seule pour laquelle les fonctions $\hat{a}$ sont toutes continues ; on considère

toujours le spectre A(A) muni de cette topologie (topologie de Gelfand).

La correspondance entre caractères et idéaux maximaux de A se matérialise alors de la manière suivante : l'idéal maximal associé au caractère χ (le noyau de χ) est l'ensemble des éléments a de A dont les transformées de Gelfand s'annulent au point χ de A(A).

Reprenons l'exemple (1) dans le cas où X est un espace compact ; il est assez facile de voir que les caractères de A sont définis par les points de X : au point x on associe le caractère χ_x tel que, pour la fonction continue bornéefsur X, élément de A, on ait $\chi_x(f) = f(x)$. On obtient ainsi tous les caractères, et cette correspondance donne un homéomorphisme entre X et A(A) qui permet d'identifier les éléments de A et leurs transformées de Gelfand.

Pour l'exemple (3), dans le cas d'un groupe discret commutatif, le spectre de A s'identifie au groupe compact dual, et la transformation de Gelfand correspond alors à la transformation de Fourier (cf. analyse HARMONIQUE).

L'ensemble des valeurs prises par la transformée de Gelfand $\hat{a}$ d'un élément a de l'algèbre normée commutative unitaire A est appelé le *spectre* de a (bien distinguer entre le spectre de l'algèbre et le spectre d'un élément de l'algèbre). Pour tout $a \in$ A, et tout $\chi \in$ A(A), on a :

$$|\hat{a}(\chi)| = |\chi(a)| \leqslant \|a\|.$$

On appelle rayon spectral de a le nombre $\|\hat{a}\|_\infty$, borne supérieure des $|\hat{a}(\chi)|$, pour χ dans A(A). L'application qui à tout élément a associe son rayon spectral est une semi-norme (car elle peut s'annuler pour a $\neq 0$) inférieure ou égale à la norme de A. On peut, à ce propos, énoncer le « théorème du rayon spectral » suivant.

Théorème. Pour tout élément a de A. on a :

$$\|\hat{a}\|_\infty = \lim_{n\to\infty} \|a^n\|^{1/n}.$$

Les algèbres semi-simples

Si la transformation de Gelfand est injective, c'est-à-dire si deux éléments distincts a et b de A ont des transformées $\hat{a}$ et $\hat{b}$ distinctes, on dit que l'algèbre normée considérée est *semi-simple*. Cela revient à dire que l'intersection des idéaux maximaux ne contient que l'élément 0.

La transformation de Gelfand permet alors d'interpréter toute algèbre normée commutative unitaire semi-simple comme une sous-algèbre de l'algèbre des fonctions continues sur un espace compact, qui est le spectre de l'algèbre donnée.

Les algèbres semi-simples jouissent de diverses propriétés spéciales : par exemple, soit A et B deux algèbres normées commutatives unitaires, B étant semi-simple ; si h est un homomorphisme algébrique de A dans B (c'est-à-dire une application telle qu'on ait :

$$h(\lambda u) = \lambda h(a),$$
$$h(a+b) = h(a) + h(b), \qquad h(ab) = h(a)\,h(b)$$

pour **tout** λ dans C et tout choix de a ct b dans A), alors h est automatiquement continu. Comme cas particulier, dans le cas où B est le corps C des nombres complexes, on retrouve la continuité des caractères.

Le calcul fonctionnel holomorphe

Soit A une algèbre normée commutative unitaire et a un élément de A; appelons o(u) le spectre de a, ensemble des nombres complexes qui sont les valeurs prises par $\hat{a}$, transformée de Gelfand de a.

Sifest une fonction continue à valeurs complexes définie sur o(a), on peut considérer la fonction composée $f \circ \hat{a}$ et se demander s'il existe un élément b dans l'algèbre tel que $f \circ \hat{a}$ soit la transformée de Gelfand de b.

Si l'algèbre A est semi-simple, il est clair qu'il peut exister au maximum un seul élément b de cette sorte. Si la fonction f est quelconque, il n'y a en général aucune raison pour qu'il existe un tel b ; mais si A est l'algèbre des fonctions continues sur un espace compact et si f est simplement supposée continue, ce sera le cas,

Supposons maintenant quef'soit définie par une série entière :

$$f(z) = \sum_{n=0} c_n z^n,$$

dont le rayon de convergence soit supérieur au rayon spectral de a. Alors la série :

$$\sum c_n a^n$$

converge dans A vers un élément b, noté $f(a)$, dont la transformée de Gelfand est précisément $\hat{b} = f \circ \hat{a}$. Cela résulte de propriétés élémentaires.

Moins simple est le théorème suivant, qui généralise de beaucoup les considérations ci-dessus :

Soit A une algèbre normée commutative unitaire semi-simple, a un élément de A et f une fonction analytique définie sur un voisinage du spectre de a. Il existe un élément b de A, et un seul, tel que $\hat{b} = f \circ \hat{a}$.

Si, en particulier, f est un polynôme :

$$f(z) = \sum_{n=0}^{N} c_n z^n,$$

b est alors l'élément :

$$\sum_{n=0}^{N} c_n a^n.$$

Si 0 n'appartient pas au spectre de a, ce qui signifie que a est inversible, et si f(z) = $1/z$, alors b est l'inverse de a.

Un cas historiquement important de ce cas particulier, dont la théorie des algèbres normées permet de donner une démonstration simple et pénétrante, est le résultat suivant :

Théorème de Wiener-Levy. Soit une fonctionfcontinue, par exemple de période 2π, et admettant pour série de Fourier :

$$\sum_{n=-\infty}^{+\infty} c_n e^{int},$$

où la série

$$\sum_{n=-\infty}^{+\infty} c_n$$

est absolument convergente. Alors, si f ne s'annule jamais, la fonction inverse $1/f$ possède une série de Fourier :

$$\sum_{n=-\infty}^{+\infty} d_n e^{int},$$

où la série :

$$\sum_{n=-\infty}^{+\infty} d_n$$

est absolument convergente.

Les algèbres normées commutatives non unitaires

Il existe un procédé standard pour associer à toute algèbre normée A une algèbre normée *unitaire* A,, telle que A soit une sous-algèbre de A,. Cc procédé, assez élémentaire, permet en principe de ramener l'étude de problèmes concernant 4 à des problèmes qui portent sur A,. Cependant, dans bien des cas, cette appïûche est insuffisante et il faut étudier directement les propriétés d'une algèbre non unitaire.

L'outil fondamental dans le cas commutatif unitaire, l'étude des idéaux maximaux, ne s'applique pas directement au cas non unitaire : il faut introduire la notion d'*idéal régulier*.

Dans une algèbre A, un idéal 1 définit une relation d'équivalence : a et b, éléments de A, sont équivalents si $a-b$ appartient à I. L'ensemble des classes d'équivalence, le quotient A/I, est muni d'une structure d'anneau (cf. ANNEAUX ET ALGÈBRES, chap. 3). On dit qu'un idéal I de l'algèbre normée commutative A est *régulier* si l'anneau quotient A/I est unitaire (remarquer que, si A est unitaire, tout idéal est régulier).

Dans une algèbre commutative unitaire, tout idéal propre est contenu dans un idéal maximal : il n'en va pas toujours de même si l'algèbre n'est pas unitaire. Mais. si l'on se borne à considérer les idéaux réguliers, on retrouve des propriétés analogues à celles qui ont été données précédemment :

a) Tout idéal régulier propre est contenu dans un idéal maximal régulier ;

b) Tout idéal maximal régulier est fermé ;

c) Tout idéal maximal régulier est le noyau d'un caractère, et d'un **seul,** et tout caractère admet pour noyau un idéal maximal régulier.

Cela définit une bijection entre l'ensemble des caractères et l'ensemble des idéaux maximaux réguliers et cela montre, d'autre part, que tout caractère est continu.

Comme pour les algèbres commutatives unitaires, on peut définir ici le spectre et la transformation de Gelfand : si A(A) est l'ensemble des idéaux maximaux réguliers de l'algèbre A, on associe à tout élément a de A sa transformée de Gelfand $\widehat{a}$, fonction définie sur A(A) de la même manière que précédemment. On munit A(A) d'une topologie d'espace localement compact, pour laquelle les transformées de Gelfand $\widehat{a}$ sont continues et tendent vers 0 à l'infini.

Cela étant, la plupart des propriétés valables pour les algèbres normées commutatives unitaires s'étendent au cas non unitaire sans modifications essentielles.

Citons, comme exemple d'algèbres de ce type, l'algèbre des fonctions continues sur un espace localement compact X qui tendent vers 0 à l'infini : ici le spectre s'identifie à X, et les éléments de l'algèbre s'identifient à leurs transformées de Gelfand. Un autre exemple est fourni par l'algèbre, pour l'opération de convolution, des fonctions intégrables sur un groupe abélien localement compact non discret, R par exemple.

3. Les algèbres normées non commutatives

L'absence de la commutativité de la multiplication interne modifie énormément, en la compliquant notablement, la théorie des algèbres normées. Faute de pouvoir ne serait-ce que l'esquisser, nous nous bornerons à indiquer deux classes d'algèbres de ce type particulièrement importantes.

Les algèbres d'opérateurs dans les espaces de Banach

Reprenons l'exemple (2) du chapitre 1 : E étant un espace de Banach. l'ensemble L(E) des applications linéaires continues de E dans E est une algèbre normée unitaire, non commutative si E est de dimension supérieure à 1. L'étude de cette algèbre est l'un des buts de l'analyse fonctionnelle.

Il est possible, en particulier, de généraliser dans ce cadre le calcul fonctionnel holomorphe. Soit par exemple T un élément de C(E) ; on appellera *spectre* de T l'ensemble o(T) des nombres complexes λ tels que $T - \lambda I_E$, où I_E est l'opérateur identique, ne soit pas inversible : cela correspond à la notion de spectre d'un élément dans une algèbre normée commutative unitaire, défini comme ensemble des valeurs prises par la transformée de Gelfand ; si f est une fonction holomorphe d'une variable complexe, définie au voisinage de o(T), on construit un autre élément de C(E), noté $f(T)$, de telle sorte que l'on ait :

$$(f+g)(T) = f(T) + g(T),$$
$$(fg)(T) = f(T) \circ g(T) ;$$

on exige de plus f(T) = T^n si f est la fonction qui à z associe z^n. À cela s'ajoutent certaines propriétés de continuité. Cette construction se fait en utilisant la formule intégrale de Cauchy (cf. FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe, chap. 5) ; en fait, bien qu'elle ait été particulièrement étudiée pour les algèbres d'opérateurs que nous considérons ici, elle est possible dans le cas le plus général et correspond au calcul fonctionnel holomorphe auquel nous avons fait allusion dans le cas des algèbres normées commutatives unitaires (où l'hypothèse supplémentaire de semi-simplicité avait pour seul but de rendre l'exposé plus concret).

RENÉ SPECTOR

Les C*-algèbres

Parmi les algèbres normées, on distingue celles dont les propriétés particulières permettent une analyse spectrale plus poussée.

On appelle *C*-algèbre* une algèbre de Banach A vérifiant les deux propriétés suivantes :

(I) elle est munie d'une *involution*, c'est-à-dire d'une application $a \rightarrow a^*$ de A dans A telle que l'on ait, quels que soient a et b dans A et λ complexe :

$$(a^*)^* = a, \quad (a+b)^* = a^* + b^*,$$
$$(\lambda a)^* = \bar{\lambda} a^*, \ (ab)^* = b^* a^*,$$

$\bar{\lambda}$ étant le nombre complexe conjugué de λ ;

(II) la norme et l'involution sont liées par la relation :

$$\|a^* a\| = \|a\|^2, \text{ quel que soit } a \text{ dans A}$$

Donnons ici quelques exemples de C*-algèbres :

(1) L'algèbre des fonctions continues sur un espace compact ;

(1') l'algèbre des fonctions continues nulles à l'infini sur un espace localement compact (dans les deux cas l'involution est l'opérateur de conjugaison).

(2) L'algèbre C(H) des opérateurs bornés sur un espace de Hilbert H (l'involution étant l'opérateur d'adjonction relatif au produit scalaire de H) ;

(2') toute sous-algèbre fermée de C(H) stable par passage à l'adjoint ;

(2") en particulier, l'algèbre CC(H) des opérateurs compacts de H, c'est-à-dire des opérateurs qui sont limite en norme d'opérateurs de rang fini.

(3) La C*-algèbre d'un groupe localement compact G : sur l'algèbre de convolution $L^1(\mu)$ (cf. ci-dessus l'exemple 3 du chapitre 1er) on construit une involution en associant à la fonction intégrable fla fonction f^* définie par : $f^*(t) = \Delta(t^{-1}) f(t^{-1})$, où A est la fonction modulaire du groupe ; on n'obtient pas ainsi une C*-algèbre (la propriété (II) de la définition n'est pas vérifiée), mais on montre qu'il existe sur

$L^1(\mu)$ une unique norme vérifiant les propriétés (I) et (II), et l'algèbre de Banach obtenue par complétion est une C*-algèbre notée habituellement C*(G).

(1) et (1') fournissent des exemples de C*-algèbre commutative : ce sont des exemples universels dans la mesure où, pour une C*-algèbre commutative, la transformation de Gelfand est un isomorphisme. On obtient ainsi le théorème de représentation.

Une C*-algèbre commutative et unitaire est naturellement isomorphe à l'algèbre des fonctions continues sur son spectre (qui est un espace compact) ;

– Une C*-algèbre commutative est naturellement isomorphe à l'algèbre des fonctions continues nulles à l'infini sur son spectre (qui est un espace localement compact).

– Cette propriété fondamentale permet de définir dans toute C*-algèbre un calcul fonctionnel continu : si a est un élément normal (i.e. tel que a* et a commutent), on peut définir sans ambiguïté l'image $f(a)$ de a par une fonction f continue à valeurs complexes sur le spectre de a.

(2) est un exemple de C*-algèbre non commutative (si H est l'espace hilbertien de dimension 2, on obtient la plus petite de celles-ci, l'algèbre des matrices 2 X 2, qui est de dimension 4).

L'exemple (2') est universel : toute C*-algèbre est isomorphe à une sous-algèbre involutive fermée d'un $\mathcal{L}(H)$ (mais il n'y a pas de manière privilégiée de la représenter ainsi).

L'algèbre des opérateurs compacts (2") joue un rôle fondamental dans toutes les théories, anciennes et nouvelles, de classification des C*-algèbres et de recherche d'invariants.

Historiquement, les algèbres d'opérateurs dans l'espace de Hilbert (C*-algèbres et algèbres de von Neumann : voir ci-dessous) ont été introduites dans les années 1930 par J. von Neumann, à la fois pour disposer d'un formalisme algébrique dans l'étude de certains problèmes de l'analyse (l'algèbre des opérateurs différentiels par exemple), et également pour interpréter mathématiquement des phénomènes spécifiques de la physique quantique, telle l'interdépendance des observations (par exemple, l'impossibilité de mesurer simultanément la position et la vitesse d'une particule est formalisée par Heisenberg comme une relation de non-commutation entre des opérateurs de l'espace hilbertien). Le formalisme abstrait que nous avons présenté est dû à I. M. Gelfand.

Rapidement, les C*-algèbres se révèlent un outil important de l'analyse harmonique, et la C*-algèbre C*(G) (exemple 3 ci-dessus) peut être considérée comme un « objet dual » du groupe localement compact G (dans le cas où G est commutatif, la transformation de Gelfand identifie C*(G) et l'algèbre C,(G) des fonctions nulles à l'infini sur le groupe dual G, ce qui est une autre manière d'écrire la dualité de Pontriaguine).

D'une manière heuristique, on peut considérer les C*-algèbres comme des « espaces localement compacts non commutatifs », considérer leur théorie comme une « topologie non commutative », leurs formes linéaires comme des « mesures non commutatives », etc. Dans ses développements récents (investigation d'invariants homotopiques et K-homologiques), leur étude tend même à s'imposer comme une « géométrie différentielle non commutative », se révélant un moyen d'investigation irremplaçable de strutures différentielles qui présentent une composante dynamique : action d'un groupe de Lie sur

une variété, et, plus généralement, toutes les structures de variété feuilletée.

Le cadre et les méthodes de la topologie algébrique ont été renouvelés par l'introduction systématique des C*-algèbres (travaux de G. Kasparov et A. Connes). Les C*-algèbres ont démontré leur aptitude à fournir et élucider des invariants topologiques pour les structures différentielles. L'effort porte aujourd'hui principalement sur la K-homologie algébrique (cf. algèbre TOPOLOGIE) et ses rapports avec la géométrie différentielle ; il peut être résumé par ses résultats les plus importants :

- le théorème de périodicité de R. Bott qui, reformulé, fournit des suites exactes de K-homologie à six termes (alors que les suites exactes d'homologie sont en principe infinies) et permet des calculs explicites ;

le théorème de l'indice de M. F. Atiyah et I. M. Singer, dans la version achevée d'A. Connes, permet de relier des invariants dynamiques d'une variété feuilletée (l'indice analytique des opérateurs pseudo-différentiels le long des feuilles, interprété comme un élément de K-homologie du fibré cotangent au feuilletage) à des invariants de nature purement algébrique (l'indice topologique, interprété comme un élément de la K-homologie de la C*-algèbre canoniquement associée au feuilletage).

Algèbres de von Neumann

Une algèbre de von Neumann est une sous-algèbre involutive de l'algèbre C(H) des opfratcurs bornés d'un espacc dc Hilbert H (cf. ci-dessus l'exemple 2') qui vérifie l'une des trois propriétés équivalentes suivantes :

a) elle contient l'opérateur identité et elle est fermée pour la topologie de la convergence simple ;

b) elle contient l'opérateur identité et elle est fermée pour la topologie de la convergence simple faible ;

c) elle est égale à son bicommutant (le commutant d'une partie P de C(H) est l'ensemble des opérateurs bornés de H qui commutent à tous les éléments de P ; le bicommutant est le commutant du commutant).

L'équivalence des propriétés a, *b* et *c* est connue comme le théorème de commutation de J. von Neumann (1929). On peut également donner une définition plus abstraite (due à J. Dixmier et S. Sakai) : une algèbre de von Neumann est une C*-algèbre qui, en tant qu'espace normé, est le dual d'un espace de Banach.

Une algèbre de von Neumann commutative s'identifie à l'algèbre des (classes de) fonctions mesurables essentiellement bornées sur un espace mesuré. Sur toute algèbre de von Neumann, le calcul fonctionnel des C*-algèbres se prolonge en un calcul fonctionnel borélien.

Pour poursuivre l'analogie du paragraphe précédent, les algèbres de von Neumann sont des « espaces mesurés non commutatifs » et leur théorie, une « théorie de la mesure non commutative » ; elle fait un usage systématique de fonctionnelles non bornées, analogues aux mesures dites σ-finies, appelées *traces* et poids : ce sont des fonctionnelles positives, densément définies, respectant les limites croissantes.

Les traces sont celles de ces fonctionnelles sous lesquelles commute toute paire d'éléments dans leur domaine. À partir d'elles, les initiateurs de la théorie, F. J. Murray et J. von Neumann, avaient classé ces algèbres en trois types : type I, ou discrètes (dont la théorie se ramène plus ou moins au cas commutatif) ; type II, ou continues et à trace (celles qui ne sont pas

discrètes mais possèdent suffisamment de traces) ; type III, ou purement infinies (celles qui ne possèdent aucune trace). Ils avaient également démontré un résultat d'unicité remarquable : celle d'une algèbre de von Neumann continue, à trace finie, à centre trivial, qui soit limite inductive d'algèbres de matrices (théorème d'unicité du facteur hyperfini de type II,, 1943).

La connaissance de la structure des algèbres de von Neumann a fait des progrès remarquables. D'abord avec la théorie de M. Tomita qui associe à **tout** poids un groupe à un paramètre d'automorphismes, le groupe modulaire, mesurant exactement son degré de non-commutativité ; ensuite avec la classification de A. Connes (dont les travaux sur les algèbres de von Neumann et les C*-algèbres ont été consacrés par une médaille Fields en 1982), fondée sur le caractère intrinsèque du groupe modulaire, qui fournit, pour le type III, des invariants affinant la typologie de Murray et von Neumann, puis généralise le théorème d'unicité du facteur hyperfini en montrant que, pour toute une catégorie d'algèbres de von Neumann (à une exception près, celles des algèbres dont le centre est trivial et qui sont limite inductive d'algèbres de matrices), il s'agit d'invariants complets, c'est-à-dire caractérisant l'algèbre à isomorphisme près.

JEAN-LUC SAUVAGEOT

Bibliographie

A. CONNES, *'Géométrie non commutative*, Interéditions, Paris, 1990 / J. DIXMIER, *Les C*-Algèbres et leurs représentations*, 2' éd. rev., Gauthier-Villars, Paris, 1969 ; *Les Algèbres d'opérateurs dans l'espace hilbertien (Algèbres de von Neumann), ibid., 2'* éd. rev. et augm., 1969 ; *C Algebras*, Elsevier Science. New York, 1977 / 1. M. GELFAND, D. A. RAIKOV & G. E. CHILOV. *Les Anneaux normés commutatifs*, trad. J.-L. et M. Verley, Gauthier-Villars, 1965 / R. G. DOUGLAS, *C-Algebra Extensions and K-Homology*, Princeton Univ. Press, 1980 / C. E. RICKART, *General Theory of Banach Algebras*, Van Nostrand. Princeton, 1960.

NORMÉS ESPACES VECTORIELS

L'analyse fonctionnelle linéaire, en tant que théorie générale, s'est créée au début du XXe siècle, autour des problèmes posés par les équations intégrales. Entre 1904 et 1906, D. Hilbert (1862-1943) est amené à étudier des développements en séries de fonctions orthogonales, ainsi que des formes quadratiques à une infinité de variables. À sa suite, F. Riesz (1880-l 956) et E. Fischer (18751959) étudient les fonctions de carré intégrdble et la convergence en moyenne quadratique, puis F. Riesz introduit les espaces L^p pour $1 < p < +\infty$ et la moyenne d'ordre p. Toutefois, ce n'est que vers 1920 que la notion d'espace normé abstrait est dégagée, principalement par S. Banach (1892-1945), et ce n'est qu'en 1929-1930 que J. von Neumann (1903-l 957) propose une présentation axiomatique des espaces de Hilbert. S. Banach, dans sa thèse de 1920 intitulée *: Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales*, écrit : « L'ouvrage présent a pour but d'établir quelques théorèmes valables pour différents champs fonctionnels, que je spécifie dans la suite. Toutefois, afin de ne pas être obligé à les démontrer isolément pour chaque-champ particulier, ce qui serait bien pénible, j'ai choisi une voie différente que voici : je considère d'une façon générale les ensembles d'éléments dont je postule certaines propriétés, j'en déduis des théorèmes et je démontre ensuite de chaque champ fonc-

tionnel particulier que les postulats adoptés sont vrais pour lui. »

Par la suite, les espaces vectoriels normés ont été étudiés de manière autonome, notamment du point de vue de leur géométrie. Parallèlement, l'obligation, en théorie des équations aux dérivées partielles par exemple, de considérer des espaces de fonctions dont la topologie n'est pas déduite d'une norme a motivé l'introduction d'une structure plus générale : celle d'espace vectoriel topologique. Toutefois, en raison de la spécificité des problèmes et des méthodes, les espaces vectoriels normés ne doivent pas être considérés comme de simples cas particuliers d'espaces vectoriels topologiques. De plus, les espaces vectoriels topologiques les plus importants peuvent être construits en un certain sens à l'aide d'espaces vectoriels normés, et bénéficient donc pour leur étude des propriétés de ces derniers. En retour, les espaces vectoriels topologiques interviennent dans l'étude des espaces normés, notamment pour tout ce qui concerne les convergences faibles.

Dans la seconde moitié du XX[e] siècle, l'évolution de la théorie est considérable, particulièrement en ce qui concerne la géométrie des espaces de Banach et ses liens avec les ensembles d'opérateurs que l'on peut définir entre les espaces étudiés.



1. Espaces vectoriels normés, espaces de Banach : définitions et premières propriétés

Dans ce qui suit, on ne considérera que des espaces vectoriels sur le corps R des nombres réels ou sur le corps C des nombres complexes. Pour éviter de préciser à chaque fois, on désignera par K ce corps de base ; pour $a \in$ K, la notation $|\alpha|$ désignera donc soit la valeur absolue de α si K = R, soit le module de a si K = C.

Soit E un espace vectoriel sur K. On appelle norme sur E une application (notée traditionnellement $x \mapsto \|x\|$; on dit aussi que $\|x\|$ est la norme de x) de E dans l'ensemble $\mathbf{R}_+$ des nombres réels positifs ou nuls qui possède les propriétés suivantes :

(1) *Condition de séparation :*

$$\|x\| = 0 \Leftrightarrow x = 0;$$

(2) *Homogénéité :*

$$\|\lambda x\| = |\lambda| \, \|x\|,$$

quels que soient $x \in$ E et $\lambda \in$ K ;

(3) *Inégalité du triangle :*

$$\|x + y\| \leqslant \|x\| + \|y\|,$$

quels que soient $x, y \in$ E.

Un espace vectoriel muni d'une norme s'appelle un *espace vectoriel normé.* Remarquons que la restriction d'une norme à un sous-espace vectoriel est une norme, appelée norme induite, sur ce sous-espace. Si la condition de séparation n'est pas satisfaite, on dit qu'on a seulement une semi-norme ; l'espace quotient de E par la relation d'équivalence :

$$x \sim y \Leftrightarrow \|x - y\| = 0$$

est alors muni de manière naturelle d'une norme, car le nombre $\|x\|$ ne dépend que de la classe de x (espace normé associé).

Tout espace vectoriel E est un espace métrique pour la distance :

$$d(x, y) = \|x - y\|,$$

déduite de la norme. On peut donc appliquer aux espaces vectoriels normés le

langage géométrique de l'analyse (boules, ouverts et fermés, convergence, etc.) introduit dans l'article espaces MÉTRIQUES. Remarquons que si d est une distance sur un espace vectoriel déduite d'une norme, elle possède la propriété suivante d'*invariance par translation* :

$$d(x+z, y+z) = d(x,y),$$

quels que soient x, y, $z \in$ E. Ainsi, les boules de centre z sont les translatées des boules centrées à l'origine 0 de l'espace vectoriel qui s'obtiennent toutes par homothétie (d'après l'homogénéité de la norme) à partir de la boule unité ouverte :

$$B(0,1) = \{x \in E ; \|x\| < 1\}$$

ou de la boule unité fermée :

$$B_f(0,1) = \{x \in E ; \|x\| \leqslant 1\}.$$

Ces boules unités sont des ensembles convexes, et on peut reconstituer la norme à partir de la boule unité par exemple ; on suppose ici, bien entendu, K = R (cf. CONVEXITÉ, chap. 4).

On dit qu'un espace vectoriel normé E est *complet, ou* encore est un *espace de Banach*, s'il est complet pour la métrique déduite de sa norme. Cela signifie ici qu'une suite (x_n) d'éléments de E est convergente si et seulement si :

$$\lim_{p,q \to \infty} \|x_p - x_q\| = 0$$

Si E est un espace vectoriel normé, on montre facilement que son complété (au sens de la théorie des espaces métriques ; cf. espaces MÉTRIQUES, chap. 3) peut être muni d'une structure d'espace de Banach qui prolonge celle de E. Ainsi, *tout espace vectoriel normé peut être plongé dans un espace de Banach dont il soit un sous-espace* dense ; *ce complété* est unique *à* un *isomorphisme d'espace vectoriel normé près.*

Il faut enfin mentionner que les espaces vectoriels normés apparaissent comme le cadre naturel de la théorie des séries et des familles sommables (cf. SÉRIES ET PRODUITS INFINIS).

Les exemples que nous donnons maintenant fournissent un premier catalogue des espaces normés les plus courants. Remarquons que lorsque ces espaces ne sont pas complets, en vertu de ce qui a été dit précédemment, on étudie leur complété afin de se ramener à un espace de Banach.

Espaces de dimension finie

Bien entendu, l'application $x \mapsto |x|$ est une norme sur K considéré comme un espace vectoriel de dimension 1 sur lui-même (et aussi d'ailleurs de C comme espace vectoriel de dimension 2 sur R).

Plus généralement, soit E un espace de dimension finie n que l'on identifie à K^n par le choix d'une base. On considère usuellement les normes :

$$\|x\|_\infty = \sup |x_i|,$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$$

pour $x = (x_1, x_2, \ldots, x_n) \in K^n$; en fait, pour tout nombre réel positif $p > 0$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

est une norme sur K^n (espaces de Minkowski). *Sur un espace de dimension finie.* on montre que *toutes les normes sont équivalentes* (cf. définition précise in chap. 3) ; cette propriété, comme on le

verra ci-dessous, est caractéristique des espaces de dimension finie, et la situation est fondamentalement différente dans ceux de dimension infinie. Les espaces de dimension *finie* sont aussi caractérisés par le fait que leur boule unité est compacte (théorème de Riesz). Mentionnons enfin que tout espace vectoriel normé de dimension finie est nécessairement complet.

La norme $\|.\|_2$ sur $\mathbf{C}^n$ est associée au produit scalaire hermitien :

$$(x\,|y) = \sum_{i=1}^{n} x_i \bar{y}_i,$$

qui munit $\mathbf{C}^n$ d'une structure hilbertienne ; de manière générale, tout espace préhilbertien E est un espace vectoriel normé si on le munit de la norme :

$$\|x\| = \sqrt{(x\,|x)},$$

en désignant par $(x\;y)$ le produit hermitien de x et y (cf. espace de HILBERT).

Norme de la convergence uniforme

Si X est un ensemble, désignons par E = $\mathcal{B}$(X, K) l'espace vectoriel des applications bornées de X dans K (rappelons que l'on a toujours K = R ou C) ; on appelle ***norme de la convergence uniforme*** la norme sur E définie par :

$$\|f\|_\infty = \sup_{t\in X}|f(t)|, \quad f \in E.$$

On montre que $\mathcal{B}$(X, K), muni de la norme de la convergence uniforme est un espace ***complet.*** Dire qu'une suite (f_n) de fonctions converge vers f pour cette norme signifie ici que :

$$\sup_{t\in X}|f(t) - f_n(t)| \to 0, n \to \infty,$$

c'est-à-dire que la suite des fonctions f_n converge uniformément vers f. Dans le cas où X est un espace topologique, on montre que le sous-espace $\mathcal{C}_\infty$(X, K) des applications ***continues*** bornées de X dans K est fermé dans $\mathcal{B}$(X, K) et, par suite (cf. espaces MÉTRIQUES, chap. 3), est aussi un espace de Banach pour la norme de la convergence uniforme.

Espaces liés à l'intégration

Soit ***[a,*** b] un intervalle fermé borné de R ; désignons par ***C([a,*** b], K) l'espace vectoriel des fonctions continues définies sur ***[a,*** b] à valeurs dans K ; pour tout nombre réel $p \geqslant 1$, on peut considérer la norme :

$$\|f\|_p = \left(\int_a^b |f(t)|^p dt\right)^{1/p},$$

appelée norme de la convergence en moyenne d'ordre ***p.*** Ces normes sont deux à deux non équivalentes, et ***C([a,*** b], K) n'est complet pour aucune d'entre elles (alors que ***C([a,*** b], K) est complet pour la norme de la convergence uniforme). Le complété de ***C([a,*** b], K) pour une telle norme (cf. espaces MÉTRIQUES, chap. 3) n'est autre que l'espace L^p ***([a, b],*** K) des classes de fonctions à valeurs dans K, de puissancepième intégrale sur ***[a,*** b] pour la mesure de Lebesgue (cf. INTÉGRATION ET MESURE, chap. 4). Pour $p = 2$ on obtient un espace de Hilbert, la norme étant associée au produit scalaire :

$$(f\,|g) = \int_a^b f(x)\overline{g(x)}\,dx.$$

Rappelons qu'une fonction mesurable définie sur ***[a,*** b] à valeurs dans K (cf. *INTÉGRATION* ET MESURE, chap. 3) est dite essentiellement bornée par M si la mesure de l'ensemble des x tels que $|f(x)| > M$ est nulle ; la ***borne supérieure essentielle,*** notée $\|f\|_\infty$, est le plus petit M réalisant la condition précédente.

L'espace L^∞ ***([a, b],*** K) des classes de fonctions essentiellement bornées sur ***[a,*** b]

à valeurs dans K est normé par $\|f\|_\infty$; c'est alors un espace de Banach.

Il existe d'autres exemples intéressants d'espaces de Banach liés à l'intégration, notamment les espaces d'Orlicz (cf. CONVEXITÉ - Fonctions convexes).

Espaces de suites

Sur l'espace l^∞ des suites bornées d'éléments de K on peut définir la norme :

$$\|u\|_\infty = \sup_{n \in \mathbf{N}} |u_n|,$$

où u est la suite de terme général u_n ; on obtient ainsi un espace de Banach. Remarquons que cet exemple peut être considéré comme un cas particulier de norme de convergence uniforme sur un espace $\mathcal{B}(X, K)$ de fonctions bornées en prenant $X = N$.

L'espace c_0 des suites d'éléments de K qui convergent vers 0 (et sont donc bornées), muni de la norme induite par la norme $\| \ \|_\infty$ de l^∞, est un espace de Banach ; c'est un sous-espace fermé de l^∞. On dispose d'un résultat analogue pour l'espace c des suites convergentes d'éléments de K.

Pour $p \geqslant 1$ on définit l'espace l^p des suites $u = (u_n)_{n \geqslant 0}$ d'éléments de K telles que $\sum_{n=0}^{\infty} |u_n|^p < +\infty$; muni de la norme :

$$\|u\| = \left(\sum_{n=0}^{\infty} |u_n|^p\right)^{1/p},$$

l^p est un espace de Banach. Dans le cas $p = 2$, on obtient un espace de Hilbert, la norme étant déduite du produit scalaire

$$(u \mid v) = \sum_{n=0}^{\infty} u_n \bar{v}_n.$$

Dans toute la suite, les espaces l^∞, c, c_0, l^p, L^p ([a, b], K) seront considérés comme normés de la façon indiquée dans les exemples D'autre part, du point de vue des notations, lorsque aucune confusion n'en résulte, on se permettra de noter C(X), L^p ([a, b]) les espaces C(X,K), L^p ([a, b], K).

Continuité d'une application linéaire

Soit E et F des espaces vectoriels normés sur K (égal à R ou C) et :

$$u : E \to F$$

une application *linéaire*, c'est-à-dire telle que :

$$u(\lambda x + \mu y) = \lambda u(x) + \mu u(y),$$

quels que soient x, $y \in E$ et λ, $\mu \in K$. Les trois conditions suivantes, apparemment de plus en plus fortes, sont en fait équivalentes :

(1) L'application u est continue au point $0 \in E$;

(2) L'application u est continue partout ;

(3) Il existe une constante M telle que :

$$\|u(x)\|_F \leqslant M\|x\|_E,$$

pour tout $x \in E$ (on exprime cette condition en disant que u est *borné*).

Ainsi, la continuité en un seul point (on se ramène à l'origine par translation) entraîne que u est uniformément continue (et même lipschitzienne, cf. espaces MÉTRIQUES, chap. 2), car on a (la linéarité est bien entendu ici essentielle) : $u(x) - u(y) = u(x - y)$, d'où :

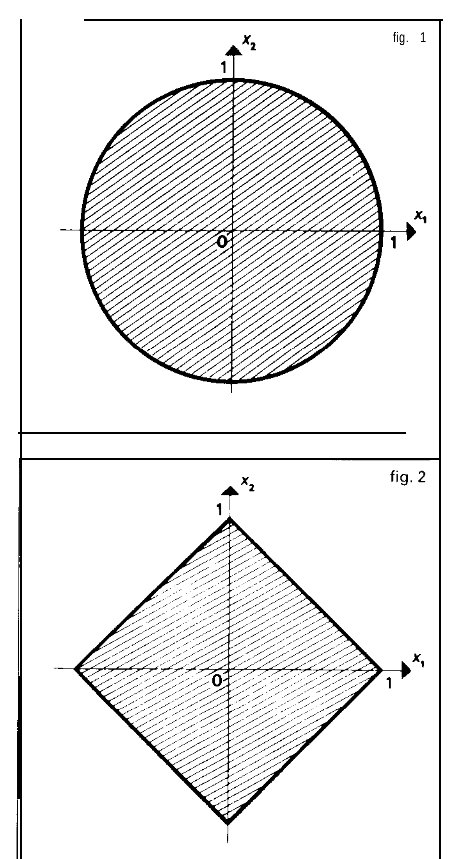
$$\|u(x) - u(y)\| \leqslant M\|x - y\|.$$

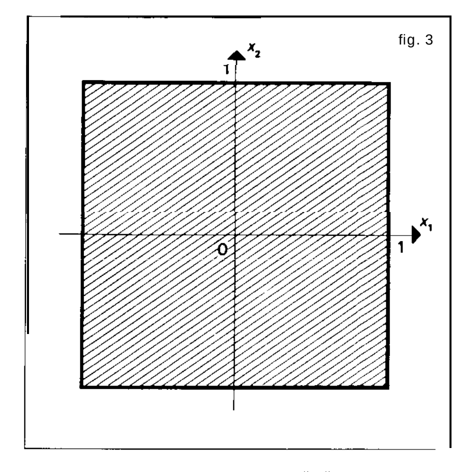
Cette importante propriété rend les applications linéaires continues redevables des résultats relatifs aux applications uniformément continues. Le théorème de

prolongement (cf. espaces MÉTRIQUES, chap. 3) donne ici : soit E un espace vectoriel normé, E' un sous-espace dense et u : E' ⟶ F une application linéaire continue de E' dans un espace de Banach F ; alors il existe un prolongement linéaire continu unique $\tilde{u}$: E ⟶ F à l'espace E tout entier (la linéarité du prolongement est évidente par continuité).

Comparaison de normes

Considérons deux normes $\|.\|_1$ et $\|\ \|_2$ sur un même espace vectoriel E et désignons par E_1 et E_2 les espaces vectoriels normés correspondants. On dit que la norme $\|.\|_1$





est plus fine que la norme $\|.\|_2$ si l'application identique de E_1 dans E_2 est continue, ce qui signifie que tout ouvert pour la norme $\|.\|_2$ est un ouvert pour la norme $\|.\|_1$.

La condition ci-dessus montre que cela équivaut à dire qu'il existe une constante $a > 0$ telle que :

$$\|x\|_2 \leqslant a\,\|x\|_1,$$

pour tout Y ∈ E. On dit que les deux normes sont *équivalentes* si elles définissent les mêmes ouverts, c'est-à-dire s'il existe des constantes strictement positives a et b telles que :

$$b\,\|x\|_1 \leqslant \|x\|_2 \leqslant a\,\|x\|_1.$$

Par exemple, si E et F sont des espaces vectoriels normés de normes respectives $\|.\|$ et $\|.\|'$, on obtient sur E X F trois normes équivalentes en prenant pour norme de l'élément $(x, y) \in$ E X F respectivement l'un des trois nombres :

$$\sup(\|x\|, \|y\|'),\quad \|x\| + \|y\|',\quad \sqrt{\|x\|^2 + \|y\|'^2}.$$

On montre que, sur un espace vectoriel de dimensionjnic, toutes les normes sont équivalentes, mais inversement ce n'est plus nécessairement le cas en dimension

infinie, comme le démontre l'exemple de l'espace $E = C([0, 1])$ muni des normes :

$$\|f\|_\infty = \sup_{0 \leqslant t \leqslant 1} |f(t)|, \quad \|f\|_1 = \int_0^1 |f(t)|\,dt ;$$

on a ici $\|f\|_1 \leqslant \|f\|_\infty$, et, par suite, la norme de la convergence uniforme est plus fine que la norme de la convergence en moyenne, mais ces deux normes ne sont pas équivalentes. Il suffit pour s'en convaincre de remarquer que la suite (f_n) d'éléments de E, définie par :

$$f_n(t) = \begin{cases} -nt + 1, & 0 \leqslant t \leqslant \frac{1}{n}, \\ 0, & \frac{1}{n} \leqslant t \leqslant 1, \end{cases}$$

converge en moyenne vers la fonction 0 mais ne converge pas uniformément, car $\|f_n\|_1 = 1/(2n)$ et $\|f_n\|_\infty = 1$.

Norme d'une application linéaire

Si E et F sont des espaces vectoriels normés, on désigne par $\mathcal{L}_c(E, F)$ l'espace vectoriel des applications linéaires *continues* de E dans F. La présence du c en indice est destinée à éviter la confusion avec l'ensemble de *toutes* les applications linéaires (continues ou pas) de E dans F que les algébristes notent (cf. algèbre LINÉAIRE ET MULTILINÉAIRE) L(E, F) ; dans la pratique, cet indice saute, car le contexte indique toujours assez clairement si on impose la continuité ou pas... Dans ce qui suit, nous ne considérerons que des applications linéaires *continues* et, le lecteur (éventuel) étant prévenu, nous désignerons par C(E, F) l'espace vectoriel des applications linéaires continues de E dans F.

On vérifie que l'application :

$$u \mapsto \|u\| = \sup_{\|x\| \leqslant 1} \|u(x)\|, \quad u \in \mathcal{L}(E, F),$$

est une norme sur l'espace vectoriel L(E, F). Le nombre $\|u\|$ s'appelle la norme *de l'application linéaire* u et admet aussi les expressions :

$$\|u\| = \sup_{\|x\| = 1} \|u(x)\| = \sup_{x \neq 0} \frac{\|u(x)\|}{\|x\|};$$

ainsi, c'est la plus petite constante M telle que l'on ait :

$$\|u(x)\| \leqslant M \|x\|,$$

pour tout $x \in E$.

On montre que l'espace vectoriel normé C(E, F) est *complet* si et seulement si F est complet. En particulier, le *dual topologique* C(E, K), qui est l'espace vectoriel des formes linéaires continues sur E, est toujours un espace de Banach ; on le note E* (ne pas confondre avec le dual algébrique).

Remarquons enfin que si E, F et G sont trois espaces vectoriels normés, si $u : E \to F$ et $v : F \to G$ sont des applications linéaires continues, alors on a :

$$\|u \circ v\| \text{ a } \|u\| \|v\|$$

en particulier, l'algèbre L(E, E) $= \mathcal{L}(E)$ des endomorphismes d'un espace vectoriel normé est une algèbre normée pour la norme introduite ci-dessus (cf. algèbres NORMÉES).

Hyperplans fermés

Soit E un espace vectoriel normé et F un sous-espace vectoriel de E. Si x et y appartiennent à l'adhérence $\overline{F}$ de F dans E, cela signifie qu'il existe des suites (x_n) et (y_n) d'éléments de F qui convergent respectivement vers x et y ; pour A, $\mu \in K$, la suite $(\lambda x_n + \mu y_n)$ d'éléments de F converge vers $\lambda x + \mu y$ qui appartient donc aussi à $\overline{F}$. Ainsi, l'adhérence d'un sous-espace vectoriel est un sous-espace vectoriel. Si F est un sous-espace de dimension *finie* de E, on montre qu'il est toujours fermé, mais, dans

les espaces de dimension infinie, il peut exister des sous-espaces distincts de leur adhérence. comme on va le voir.

Rappelons (cf. algèbre LINÉAIRE ET MULTILINÉAIRE, chap. 4) qu'on appelle *hyperplan* d'un espace vectoriel tout sous-espace strict maximal, c'est-à-dire de codimension 1 ; si H est un hyperplan de E, il existe une forme linéaire $u : E \to K$, unique à un scalaire près, telle que H soit le noyau de u (on dit que $u(x) = 0$ est l'équation de l'hyperplan). Supposons E normé et soit H un hyperplan ; l'adhérence $\overline{H}$ est un sous-espace vectoriel de E qui contient H, et par suite, d'après la maximalité de H, on a soit $H = \overline{H}$, c'est-à-dire que l'hyperplan H est *fermé*, soit $\overline{H} = E$, c'est-à-dire que l'hyperplan est ***dense*** dans E. On montre facilement que, avec les notations données ci-dessus, I'hyperplan H est fermé si et seulement si la forme linéaire u est continue.

La notion d'hyperplan partout dense étant peu intuitive, donnons un exemple simple de cette situation. Soit E l'espace vectoriel des polynômes à coefficients réels, muni de la norme de la convergence uniforme sur [0, 1], c'est-à-dire :

$$\|P\| = \sup_{0 \leqslant t \leqslant 1} |P(t)|, \quad P \in E;$$

la forme linéaire u définie par :

$$u(P) = P(2)$$

n'est pas continue, car la suite (P_n), définie par $P_n(t) = (t/2)^n$, converge vers 0 dans E ; en effet, $\|P_n\| = (1/2)^n$), alors que $u(P_n) = 1$ ne converge pas vers 0. Il en résulte que l'espace vectoriel des polynômes P qui admettent le nombre 2 pour racine forme un hyperplan partout dense de E.

Isomorphismes. isométries

Une application linéaire bijective u d'un espace normé E sur un espace normé F telle que u et u^{-1} soient continues est un *isomorphisme* de E sur F ; deux espaces normés E et F sont *isomorphes* s'il existe un isomorphisme de E sur F ; du point de vue topologique, les espaces E et F sont homéomorphes (cf. TOPOLOGIE GÉNÉRALE, chap. 1). Compte tenu de ce qui a été dit sur la continuité des applications linéaires, pour qu'une application linéaire surjective de E sur F soit un isomorphisme il faut et il suffit qu'il existe deux constantes $C_1 > 0$ et $C_2 > 0$ telles que pour tout élément x de E :

$$C_1 \|x\|_E \leqslant \|u(x)\|_F \leqslant C_2 \|x\|_E.$$

(Remarquons que l'injectivité est conséquence de l'inégalité $C_1 \|x\|_E \leqslant \|u(x)\|_F$ et de la linéarité de u, si bien que si u n'est pas surjective on peut tout de même dire que u est un isomorphisme de E sur u(E).)

Une application linéaire bijective u d'un espace normé E sur un espace normé F telle que pour tout x de E $\|u(x)\|_F = \|x\|_E$ est une *isométrie*, ou encore un *normisomorphisme* de E sur F ; s'il existe une isométrie de E sur F, les espaces E et F sont dits *isométriques* ou encore *normisomorphes*.

Tous les espaces de Banach de même dimension finie n sur K sont isomorphes : en revanche. ils ne sont pas tous isométriques comme le montre la considération des normes $\| \ \|_2$ et $\| \ \|_\infty$ par exemple. Si $1 \leqslant p < q < +\infty$, aucun sous-espace fermé de dimension infinie de l_p n'est isomorphe à un sous-espace de l_q ; aucun sous-espace fermé de c_0 n'est isomorphe à un sous-espace de l_p. K_1 et K_2 étant deux espaces compacts, $C(K_1, R)$ et $C(K_2, R)$ sont isométriques si et seulement si K_1 et K_2 sont homéomorphes ; C ([0,1], R) et C ([0,1] × [0,1], R) ne sont donc pas isométriques ; on peut montrer cependant qu'ils sont isomorphes.

2. Les théorèmes généraux de base

Entre 1920 et 1930, S. Banach, H. Hahn, H. Steinhaus élaborent les théorèmes généraux de base de la théorie.

Théorème de Hahn-Banach

Il existe diverses versions de ce théorème ; nous donnons ici une version analytique valide dans les deux cas : K = R ou K = C. Nous renvoyons à l'article CONVEXITÉ pour une forme géométrique de ce résultat.

Soit E un espace vectoriel sur K, p une semi-norme sur E (cf. chap. 1) et f une forme linéaire sur un sous-espace F de E qui pour tout x de F vérifie $|f(x)| \leqslant p(x)$. II existe alors une forme linéaire g sur E qui prolongefet qui vérifie $g(x) \leqslant p(x)$ pour tout x de E.

Les quatre théorèmes qui suivent reposent de manière essentielle sur la propriété de Baire des espaces métriques complets (cf. espaces MÉTRIQUES, chap. 4).

Théorème de l'application ouverte

Soit E et F deux espaces de Banach et u une application linéaire continue surjective de E sur F. L'image par u de tout ouvert de E (cf. TOPOLOGIE GÉNÉRALE, chap. 1) est alors un ouvert de F.

On déduit immédiatement de ce théorème que si de plus u est injective alors u est un isomorphisme de l'espace de Banach E sur l'espace de Banach F. En particulier, lorsqu'un espace vectoriel E est muni de deux normes qui en font toutes deux un espace de Banach, il suffit de montrer que ces normes se comparent pour en conclure qu'elles sont équivalentes.

Théorème du graphe fermé

Soit E et F deux espaces de Banach. Pour qu'une application linéaire u de E dans F soit continue, il faut et il suffit que son graphe soit fermé dans l'espace produit E X F.

Théorème d'équicontinuité de Banach

Soit $(T_i)_{i \in I}$ une famille d'applications linéaires continues d'un espace de Banach B dans un espace vectoriel normé F. On suppose que, pour tout élément x de B, $\sup_{i \in I} \| T_i(x) \| < +\infty$; alors $\sup_{i \in I} \| T_i \| < +\infty$.

Théorème de Banach-Steinhaus

Soit E et F deux espaces de Banach et $(T_n)_{n \in \mathbb{N}}$ une suite d'applications linéaires continues de E dans F. Alors $\lim_{n \to \infty} T_n(x)$ existe pour tout x élément de E si et seulement si $\lim_{n \to \infty} T_n(x)$ existe pour tout x d'un sous-ensemble dense de E et $\sup_n \| T_n(x) \| < +\infty$ pour tout x élément de E. Quand la limite T(x) existe pour tout élément x de E, l'application T est linéaire continue et $\| T \| \leqslant \liminf_{n \to \infty} \| T_n \|$.

3. La décomposition des espaces de Banach

Produits d'espaces de Banach

E et F étant deux espaces de Banach, la somme directe $E \oplus F$ (cf. algèbre LINÉAIRE ET MULTILINÉAIRE, chap. 2) peut être munie d'une structure d'espace de Banach dont la topologie associée soit la topologie produit de celle de E par celle de F (cf. TOPOLOGIE GÉNÉRALE, chap. 1). Il y a en fait plusieurs normes qui réalisent cette condition, les plus utilisées étant $\| (x, y) \|_p = (\| x \|_E^p + \| y \|_F^p)^{1/p}$, où $1 \leqslant p < +\infty$ et $\| (x, y) \|_\infty = \max (\| x \|_E, \| y \|_F)$. Évidemment, ces normes sont équivalentes et les espaces de Banach obtenus sont isomorphes.

La complémentation

Soit $E \oplus F$ une décomposition en somme directe algébrique de l'espace de Banach X. E et F étant munis des topologies induites par celle de X, et $E \oplus F$ de la topologie produit, nous dirons que $E \oplus F$ est une décomposition en somme directe topologique si l'application

$$\begin{aligned} \varphi : E \oplus F &\to X \\ (x, y) &\mapsto x + y \end{aligned}$$

est un homéomorphisme.

En utilisant le théorème de l'application ouverte, on montre que pour qu'une décomposition en somme directe $E \oplus F$ de l'espace de Banach X soit topologique il faut et il suffit que E et F soient des sous-espaces fermés de X.

Le problème de la complémentation qui se pose alors est de savoir si, étant donné un sous-espace vectoriel fermé E d'un espace de Banach X, il existe un supplémentaire topologique de E dans X, c'est-à-dire un sous-espace F de X tel que $E \oplus F$ soit une décomposition en somme directe topologique de X ; on dira dans ce cas que E est complémenté dans X. On montre que pour qu'un sous-espace fermé E de X soit complémenté dans X il faut et il suffit qu'il existe une projection continue P de X sur E ; alors $E \oplus (I - P)(X)$, où I est l'application identique, est une décomposition en somme directe topologique de X. Il n'est pas vrai en général que tous les sous-espaces fermés d'un espace de Banach soient complémentés : par exemple, c_0 n'est pas complémenté dans l^∞. Toutefois, la propriété indiquée est réalisée dans les espaces de dimension finie et dans les espaces de Hilbert (cf. espace de HILBERT, chap. 3) et elle caractérise ces espaces ; plus précisément

a) Soit E un espace de Banach dans lequel il existe pour tout sous-espace F fermé de E une projection continue de norme 1 de E sur F ; alors E est isométrique à un espace de Hilbert.

b) Soit E un espace de Banach dans lequel il existe pour tout sous-espace F fermé de E une projection continue de E sur F ; alors E est isomorphe à un espace de Hilbert.

Bases de Schauder

Soit $(x_i)_{i \in \mathbf{N}}$ une suite d'éléments d'un espace de Banach E telle que tout élément x de E se décompose de manière unique sous la forme $x = \sum_{i=0} x_i^*(x) x_i$, où les $x_i^*(x)$ sont des éléments de K (qui dépendent évidemment de x). Dans ces conditions, les applications :

$$\begin{aligned} x_i^* : E &\to K \\ x &\mapsto x_i^*(x) \end{aligned}$$

sont des formes linéaires continues, c'est-à-dire des éléments de E*. On dit alors que la suite $(x_i)_{i \in \mathbf{N}}$ est une *base de Schauder* de E. Dans les espaces c_0, l_p $(1 \leqslant p < +\infty)$, la suite $(e_i)_{i \in \mathbf{N}}$, où $e_i = (\delta_{i,j})_{j \in \mathbf{N}}$ ($\delta_{ij} = 0$ si $i \neq j$, $\delta_{ij} = 1$ si $i = j$) est une base de Schauder. Dans l'espace de Banach C[0,1] muni de la norme de la convergence uniforme, la suite de fonctions définie par :

$$f_0(t) = 1, f_1(t) = t,$$

$$f_{2^k+r}(t) = \begin{cases} 0 & \text{si } t \notin \left]\frac{2r-2}{2^{k+1}}, \frac{2r}{2^{k+1}}\right[\\ 1 & \text{si } t = \frac{2r-1}{2^{k+1}} \\ \text{affine sur} & \left[\frac{2r-2}{2^{k+1}}, \frac{2r-1}{2^{k+1}}\right] \text{ et } \left[\frac{2r-1}{2^{k+1}}, \frac{2r}{2^{k+1}}\right] \end{cases}$$

$(r = 1, 2, \ldots, 2^k ;\ k = 0, 1, \ldots)$ constitue une base de Schauder.

Dans un espace de Banach E muni d'une base de Schauder, les combinaisons linéaires finies à coefficients rationnels d'éléments de la base forment une famille dénombrable dense dans E : l'espace E est séparable. La plupart des espaces de Banach séparables que l'on rencontre sont munis de bases de Schauder ; on peut néanmoins construire des espaces de Banach séparables qui n'en possèdent pas.

4. Les propriétés d'approximation

On supposera désormais que K = R.

Une application linéaire d'un espace vectoriel E dans un espace vectoriel F est dite de rang fini si son image est un sous-espace de dimension finie de F. X et Y étant deux espaces de Banach et f une application linéaire continue de rang fini de X dans Y, il est clair, d'après le théorème de Riesz (cf. chap. 1), que l'adhérence dans Y de l'image parfde la boule unité fermée de X est une partie compacte de Y, c'est-a-dire quefest un opérateur compact. Comme on sait d'autre part que dans $\mathcal{L}_c$ (X, Y) muni de la norme des applications linéaires continues une limite d'opérateurs compacts est un opérateur compact, la question qui se pose est de savoir si, pour des espaces de Banach X et Y arbitraires, tout opérateur compact de X dans Y est limite dans $\mathcal{L}_c$ (X, Y) muni de la norme indiquée d'une suite d'opérateurs de rang fini. Ce problème, dit problème de l'approximation. n'a été résolu par la négative qu'en 1973 par P. Enflo ; il a donné lieu a l'étude de divers énoncés équivalents et à la mise en place de quelques propriétés voisines extrêmement importantes (ces travaux sont essentiellement dus à A. Grothendieck). On dit que l'espace de Banach X possède la propriété d'approximation (la A.P.) si, pour tout compact Q de X et tout $\varepsilon > 0$, il existe une application linéaire continue T de rang fini telle que pour tout x de Q on ait $\| T_x - x \| \leqslant \varepsilon$. Soit λ un réel $\geqslant 1$, si on impose à T la condition supplémentaire d'être de norme inférieure à λ, on dit alors que X a la A-propriété d'approximation (λ-A.P.). Lorsqu'il existe un réel A tel que l'espace de Banach X ait la A-A.P., on dit que X a la propriété d'approximation bornée (B.A.P.).Enfin,siXala 1-A.P.,onditqu'il a la propriété d'approximation métrique. On montre que l'espace de Banach X a la A.P. si et seulement si, pour tout espace de Banach Y, tout opérateur compact de X dans Y est limite dans $\mathcal{L}_c$ (X, Y) muni de sa norme usuelle d'une suite d'applications linéaires continues de rang fini.

On sait qu'il existe des espaces de Banach qui n'ont pas la A.P. et qu'il existe des espaces de Banach qui ont la A.P. mais n'ont pas la B.A.P. Soit un espace de Banach X ayant une base de Schauder $(x_i)_{i\in\mathbf{N}}$, la considération des opérateurs de rang fini T,, défini par T,(x) $= \sum_{i=0}^{n} x_i^*(x)\, x_i$ montre que X a la B.A.P. ; on ne sait pas par contre si tout espace de Banach séparable ayant la B.A.P. possède une base de Schauder.

5. Intégration des fonctions à valeurs vectorielles. Mesures à valeurs vectorielles

L'intégration des fonctions à valeurs vectorielles et les mesures à valeurs vectorielles sont des outils intéressants qui permettent en particulier grâce à des théorèmes de représentation de mieux étudier certaines propriétés géométriques des espaces de Banach.

Intégration des fonctions à valeurs vectorielles

$(\Omega, \mathcal{T}, \mu)$ est un espace mesuré par une mesure positive finie μ ; X est un espace de Banach et $\mathcal{B}_X$ est la tribu borélienne de X (cf. INTÉGRATION ET MESURE). Une application f de Ω dans X est dite *fortement mesurable* si c'est une application mesurable (c'est-à-dire si l'image réciproque par f de tout élément de $\mathcal{B}_X$ est un élément de $\mathcal{T}$) et s'il existe un sous-espace fermé séparable X_0 de X et un élément Ω_0 de $\mathcal{T}$ de mesure nulle tels que $f(\Omega \quad \Omega_0) \subset X_0$.

Une application f de Ω dans X est dite simple si elle est mesurable et si son image est un sous-ensemble fini de X.

Une fonction simple non nulle s'écrit alors de manière unique sous la forme :

$$f = \sum_{i=1}^{n} \chi_{A_i} x_i,$$

où les x_i sont des éléments non nuls deux à deux distincts de X, où les A, sont des éléments non vides deux à deux disjoints de la tribu $\mathcal{T}$ et où χ_A est la fonction caractéristique de l'ensemble A,. On peut alors définir l'intégrale de la fonction simple $f = \sum_{i=1}^{n} \chi_{A_i} x_i$ par rapport à la mesure μ en posant :

$$\int f \, d\mu = \sum_{i=1}^{n} \mu(A_i) x_i$$

(on attribuera à la fonction nulle l'intégrale 0).

Toute fonction f de Ω dans X, fortement mesurable, est limite presque partout d'une suite de fonctions simples ; cela nous suggère de définir l'intégrale de certaines fonctions fortement mesurables grâce à une approximation par des fonctions simples ; cette démarche est possible grâce au lemme suivant :

Lemme. Soit $(f_n^1)_n$ et $(f_m^2)_m$ deux suites de fonctions simples qui convergent presque partout vers la même fonction simple f et telles que, pour $i = 1$ et 2, $\lim_{\substack{n\to\infty \\ m\to\infty}} \int_\Omega \|f_n^i(\omega) - f_m^i(\omega)\| \, d\mu = 0$. Alors, les limites $\lim_{n\to\infty} \int_A f_n^i(\omega) \, d\mu$ ($i = 1$ et 2) existent pour tout A élément de $\mathcal{T}$ (et même uniformément par rapport à A) et sont égales.

On peut donner alors la définition suivante :

Définition. Une fonction fortement mesurable f, de Ω dans X, est dite *intégrable au sens de Bochner* (ou B-intégrable) s'il existe une suite $(f_n)_n$ de fonctions simples qui converge vers f presque partout et telle que :

$$\lim_{\substack{n\to\infty \\ m\to\infty}} \int_\Omega \|f_n(\omega) - f_m(\omega)\| \, d\mu = 0.$$

Par définition, on pose alors :

$$\int_\Omega f(\omega) \, d\mu = \lim_{n\to\infty} \int_\Omega f_n(\omega) \, d\mu.$$

Théorème. Une fonction fortement mesurable f de Ω dans X est B-intégrable si et seulement si $\int_\Omega \|f(\omega)\| \, d\mu < +\infty$; dans ce cas, on a de plus l'inégalité :

$$\left\| \int_\Omega f(\omega) \, d\mu \right\| \leqslant \int_\Omega \|f(\omega)\| \, d\mu.$$

De la même façon que pour les fonctions à valeurs réelles ou complexes, on définit l'espace $L^1(\Omega, \mathcal{T}, \mu, X)$ des classes de fonctions B-intégrables de Ω dans E (cf. INTÉGRATION ET MESURE, chap. 4), ainsi que les espaces $L^p(\Omega, \mathcal{T}, \mu, X)$, où $1 \leqslant p \leqslant +\infty$.

Mesures à valeurs vectorielles

Soit $(\Omega, \mathcal{T})$ un espace mesurable et X un espace de Banach. Une application F

de $\mathcal{T}$ dans X est une mesure si elle est τ-additive : pour toute suite (A,), d'éléments deux à deux disjoints de $\mathcal{T}$:

$$F\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} F(A_i),$$

et si $F(\emptyset) = 0$.

Notons F l'application de $\mathcal{T}$ dans $\mathbf{R}_+$ définie par :

$$|F|(A) = \sup_{\Pi \in P} \sum_{B \in \Pi} \|F(B)\|,$$

où P est l'ensemble de toutes les partitions finies de A en éléments de $\mathcal{T}$; F est une mesure positive ; quand F est une mesure finie, on dit que la mesure F est à variation bornée. Si maintenant μ est une mesure positive finie sur l'espace mesurable (E, $\mathcal{T}$) et si F est une mesure définie sur (E, $\mathcal{T}$) à valeurs dans X, on dira que F est p-continue lorsque $\lim_{\mu(A)\to 0} F(A) = 0$.

On peut montrer que, pour que F soit u-continue, il faut et il suffit que, pour tout élément A de $\mathcal{T}$ tel que $\mu(A) = 0$, on ait aussi $F(A) = 0$.

La propriété de Radon-Nikodym

Théorème. Soitfun élément de L^1(E, $\mathcal{T}$, μ, X) ; alors la fonction F de $\mathcal{T}$ dans X définie par :

$$F(A) = \int_A f(\omega)\, d\mu$$

est une mesure à variation bornée et :

$$F(A) = \int_A \|f(\omega)\|\, d\mu.$$

Contrairement à ce qui se passe dans le cas de mesures à valeurs réelles ou complexes, la réciproque de ce théorème est fausse. On est amené à introduire la définition qui suit.

Définition. On dit qu'un espace de Banach X a la propriété de Radon-Nikodym si, pour tout espace mesuré (Ω, $\mathcal{T}$, μ) par une mesure positive finie μ, et pour toute mesure à variation bornée F de $\mathcal{T}$ dans X y-continue, il existe un élément f de L^1 (E, $\mathcal{T}$, μ, X) tel que pour tout A élément de $\mathcal{T}$ on ait :

$$F(A) = \int_A f(\omega)\, d\mu.$$

Cette propriété a des applications intéressantes du point de vue des opérateurs :

Définition. Soit (Ω, $\mathcal{T}$, μ) un espace mesuré par une mesure positive finie μ. On dit qu'un opérateur T de L^1 (Ω, $\mathcal{T}$, μ) dans un espace de Banach X admet une représentation de Riesz s'il existe un élément g de L^∞ (Ω, $\mathcal{T}$, μ, X) tel que pour tout élément f de L^1 (Ω, $\mathcal{T}$, μ) :

$$T(f) = \int_\Omega f(\omega)\, g(\omega)\, d\mu.$$

Théorème. Un espace de Banach X a la propriété de Radon-Nikodym si et seulement si, pour tout espace mesuré (Ω, $\mathcal{T}$, μ) par une mesure positive finie μ, tout opérateur T de L^1 (Ω, $\mathcal{T}$, μ) dans X admet une représentation de Riesz.

La propriété de Radon-Nikodym a d'autre part un aspect géométrique très important :

Définition. Un sous-ensemble borné B d'un espace de Banach X est dit dentable si, pour tout $\varepsilon > 0$, il existe un élément x de B qui n'appartient pas à l'enveloppe convexe fermée de B $-$ B(x, ε), où B(x, ε) est la boule ouverte de centre x et de rayon ε.

Théorème. Un espace de Banach X possède la propriété de Radon-Nikodym si et seulement si tout sous-ensemble borné de X est dentable.

ROBERT ROLLAND et JEAN-LUC VERLEY

Bibliographie

L. CHAMBADAL & J.-L. OVAERT, *Cours de mathématiques*, t. I : *Notions fondamentales d'algèbre et d'analyse*, Gauthier-Villars, Paris, 1966 / M. DAY, *Normed Linear Spaces*, Spinger, Berlin-Göttingen-Heidelberg. 1962 / J. DIEUDONNÉ. *Fondements de l'analyse moderne*, Gauthier-Villars. 1965 ; *Éléments d'analyse*, t. II, Gauthier-Villars. 1968 / M. KLINE, *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Press, New York. 1972 / A. KOLMOGOROV & S. FOMINE, *Éléments de la théorie des fonctions et de l'analyse fonctionnelle*, trad. M. Dragnev, MIR, Moscou. 2e éd., 1977 / G. KÖTHE, *Topological Vector Spaces*, t. I, Springer, Berlin-Heidelberg-New York. 1969 / J. LINDENSTRAUSS & L. TZAFRIRI, *Classical Banach Spaces*, t. I : *Sequence Spaces*, *ibid.*, 1977.

NUMÉRATION

Le problème de la numération est celui de la désignation des nombres. Les nombres sont définis de manière intrinsèque, indépendamment de leur nom, et la façon de les désigner dépend du langage, du « code » choisi. Pour comprendre en quoi consiste la numération, il est important d'abord de savoir distinguer un nombre de ses représentations dans divers « systèmes de numération ». Nous ne rappelons d'abord ici que les notions élémentaires concernant les nombres entiers naturels.



Les entiers naturels

Bijections

Une application *f* d'un ensemble A sur un ensemble B est dite une *bijection* lorsque : tout élément de B est l'image parfd'un Elément de A (surjection) ; deux éléments distincts de A ont toujours pour images par *f* deux éléments distincts de B (injection).

Lorsqu'il existe une bijection de A sur B, il en existe aussi une de B sur A, et on dit que A et B ont *autant* d'éléments. C'est une notion très simple, car on peut voir si deux ensembles ont autant d'éléments sans compter ces éléments.

Ainsi, les bergers de l'Antiquité utilisaient des cailloux (d'où le nom de « calcul ») pour faire rentrer le soir *autant* de moutons qu'ils en avaient fait sortir le matin ; de même, lorsqu'on voit de nombreux couples danser sur une scène, malgré l'animation et sans compter, on sait immédiatement qu'il y a *autant* d'hommes que de femmes ; remarquons enfin que, dès l'école maternelle, les enfants savent qu'ils ont *autant* de doigts à une main qu'à l'autre, aux mains qu'aux pieds, qu'il y a autant de tasses que de soucoupes, etc., et cela parce qu'ils savent réaliser les bijections correspondantes.

Cardinaux

Plusieurs ensembles d'objets étant donnés, on peut opérer un classement en rangeant dans une même « classe » les ensembles ayant *autant* d'éléments. Les ensembles d'une même classe sont dits « équipotents ». Ces exercices présentent l'inconvénient de ne porter que sur des ensembles finis, mais permettent de bien mettre en évidence la notion d'équipotence entre ensembles.

L'équipotence entre ensembles est réflexive, symétrique et transitive, mais on peut remarquer que l'on commet un abus de langage lorsqu'on dit que c'est une « relation d'équivalence ». En effet, les relations sont définies seulement sur des ensembles, or l'équipotence est définie sur la « collection de tous les ensembles », de

même que l'appartenance ou l'égalité des ensembles.

Les cardinaux peuvent être considérés comme les « classes d'équivalence » déterminées par cette « pseudo-relation » sur la collection de tous les ensembles : les ensembles d'une même classe ont donc en commun la propriété d'avoir « même cardinal ».

Nombres entiers

Aspect « cardinal »

On peut être tenté de définir le « nombre d'éléments » d'un ensemble comme la propriété commune à tous les ensembles qui ont même cardinal que lui ; dans ce cas, « nombre » et « cardinal » seraient synonymes. Mais, en réalité, seuls les cardinaux *finis* sont des nombres entiers.

On peut prendre comme définition : Un ensemble est fini si et seulement s'il n'est équipotent à aucune partie stricte de lui-même.

Aspect « ordinal »

Si l'on construit une suite d'ensembles dont le premier est vide et tels que, à partir du deuxième (auquel on donnera le numéro un), chacun s'obtient en recopiant le précédent et en lui adjoignant un objet et un seul, c'est-à-dire que chaque ensemble a exactement « un objet de plus » que le précédent, ce qu'on matérialise ainsi n'est autre que la construction des « ordinaux finis » par :

$$\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \ldots$$

qui sont respectivement de cardinal :

$$0, 1, 2, 3,$$

Or deux ordinaux finis de même cardinal sont isomorphes ; on peut sans danger les confondre et les identifier à leur cardinal ; mais, selon les situations, c'est l'aspect ordinal ou l'aspect cardinal du nombre naturel considéré qui intervient.

Numération des entiers naturels

L'ensemble des entiers naturels étant construit, la question se pose de « nommer » ces nombres oralement et par écrit, 11 apparaît vite qu'il n'est pas possible d'inventer un nom pour chaque nombre indépendamment des précédents ; il est encore moins possible de lui trouver un symbole pour l'écriture. Chaque civilisation s'est donc donné un « alphabet » particulier et des règles de formation pour les « mots », au sens de « combinaisons de symboles ».

Le système adopté par la civilisation occidentale utilise actuellement les symboles :

« 0 », « 1 », « 2 », « 3 », « 4 »,
« 5 », « 6 », « 7 », « 8 », « 9 »

qui constituent l'alphabet à partir duquel on écrit les nombres en appliquant le principe dit de « numération de position » avec une base constante.

Numération de position à base constante

Soit B un entier naturel fixe, dit « base » : une unité de chaque ordre vaut B unités de l'ordre précédent.

Par suite de l'unicité du quotient et du reste dans la division euclidienne (cf. DIVISIBILITÉ, chap. 1), tout entier naturel a peut s'écrire d'une manière et d'une seule sous la forme :

$$a = a_0 + a_1\mathrm{B} + a_2\mathrm{B}^2 + \ldots + a_n\mathrm{B}^n,$$

où les a_0, a,, ... a, sont des entiers naturels strictement inférieurs à B et où a, est non nul.

La numération de position revient à représenter le nombre en écrivant seule-

ment les coefficients de ce polynôme (mais *tous* les coefficients nuls ou non, de manière que leur place soit définie sans ambiguïté), donc à désigner le nombre précédent par :

$$\overline{a_n a_{n-1} \ldots\ldots\ldots a_2 a_1 a_0}^{B},$$

ou, plus généralement, lorsque aucune confusion n'est possible, en omettant l'indication de la base, par :

$$\overline{a_n a_{n-1} \ldots\ldots\ldots a_2 a_1 a_0}$$

et même sans surlignage par :

$$a_n a_{n-1} \cdot \quad a_2 a_1 a_0.$$

Ainsi. le nombre « neuf » s'écrit :

$$\overline{9}^{(\text{dix})} \text{ ou } \overline{1001}^{(\text{deux})} \text{ ou } \overline{100}^{(\text{trois})}, \text{ etc.}$$

Une erreur est à éviter : il faut se garder de lire « mille un » pour 100 1 $^{(\text{deux})}$; on doit lire la suite des chiffres écrits de gauche à droite dès que le nombre est écrit dans une base différente de dix. Il serait également maladroit d'écrire la base en chiffres, car on ne saurait pas de quel nombre il s'agit (sauf lorsque l'on convient que les bases sont toujours exprimées dans la base dix, par exemple).

Le *système décimal* est le système de numération de position où la base est dix, c'est-à-dire que les unités du deuxième ordre (les « dizaines ») valent dix unités du premier ordre, les unités du troisième ordre (les « centaines ») valent dix unités du deuxième ordre, etc. Prenons, par exemple, 8 345 :

$$8345 = 5 + 4 \times 10 + 3 \times 10^2 + 8 \times 10^3.$$

Le *système binaire* est le système de numération de position où la base est deux : l'alphabet est composé des deux seuls chiffres 0 et 1. Ce système est très utilisé, en informatique par exemple, car les machines à deux états peuvent réaliser une représentation des nombres entiers par leur désignation binaire, les deux états de la machine étant, dans le code, la traduction du 0 et du 1. Ainsi, « neuf » peut être codé par un top suivi de deux blancs puis d'un autre top.

Lorsque la base est supérieure à dix, il est nécessaire d'adjoindre aux chiffres habituels de nouveaux symboles, Par exemple, en base douze, on utilisera :

« 0 », « 1 », « 2 », « 3 », « 4 », « 5 »,
« 6 », « 7 », « 8 », « 9 », « α », « β ».

Numération de position à base non constante

On peut voir que, dans de nombreuses civilisations, le système de numération est un système positionnel à base non constante : il est analogue au système défini plus haut, mais les unités des divers ordres ne sont pas toutes les puissances de l'unité du premier ordre. Les unités de chaque ordre étant définies, tout nombre naturel s'écrit encore d'une manière et d'une seule dans le système déterminé par ces unités en opérant des divisions euclidiennes successives comme dans les cas à base constante.

Comparaison de deux nombres et opérations

Deux nombres écrits dans le même système de numération de position peuvent être comparés : on a vu qu'un même nombre ne peut s'écrire que d'une seule manière dans un système donné. Soit deux nombres a et b :

$$a = \overline{a_n\ a_{n-1}\ a_0}, \quad b = \overline{b_m\ b_{m-1}\ b_0}\ ;$$

si $m < n$, alors $b < a$; si $m > n$, alors $b > a$; si $m = n$, alors, ou bien, si $a_n \neq b_n$, a et b sont dans le même ordre que a_n et

b_n, ou bien, si a,, = b_n, a et b sont dans le même ordre que a_i et b_i, l'entier i étant le plus grand entier p tel que $a_p \neq b_p$.

Pour les opérations, le système de numération a des implications sur les techniques opératoires (retenues) : la désignation du résultat d'une opération sur les entiers naturels est fonction de la désignation de ces nombres.

Apprentissage de la numération

On peut présenter, dès l'école primaire, des situations mettant en lumière les principes de numération que nous venons d'énoncer.

Citons d'abord des numérations à *buse non constante* :
- dans de nombreux jeux, les enfants comptent les points gagnés en utilisant des jetons tels que, par exemple, cinq ronds valent un carré, deux carrés valent un rectangle, etc. ;
- utilisation des pièces de monnaie courantes (centimes, sous, francs) ;
- décompte des voix obtenues à des élections en dessinant des blocs de cinq traits : par exemple, ❑ i5 |_| donne treize ;

calendrier et mesure du temps.

Les exemples d'enseignement scolaire de la numération *à base constante* sont évidemment nombreux :
- le boulier traditionnel, très utilisé encore actuellement dans certains pays pour l'apprentissage des opérations ;

les exercices de groupement par paquets (trois billes dans un sac, trois sacs dans une boîte. trois boîtes dans une caissette, etc.) ;
- le solfège : dès huit ans, les enfants savent qu'une ronde vaut deux blanches, une blanche vaut deux noires, une noire vaut deux croches, une croche vaut deux doubles croches... ; ils utilisent donc ici la « numération binaire » ;
- le matériel pédagogique : les « blocs multibases » utilisés dans l'enseignement primaire sont des ensembles de petits cubes, de barres, de plaques carrées et de grands cubes ; pour compter en base trois, par exemple, on utilise des petits cubes, des barres formées de trois petits cubes accolés, des plaques formées de trois barres et des cubes formés de trois plaques ; les enfants, pour compter le nombre d'éléments d'un ensemble d'objets, ont, d'abord, à prendre « autant » de petits cubes qu'il y a d'objets (en établissant une bijection), puis ils les regroupent, remplacent chaque ensemble de trois petits cubes par une barre, puis chaque ensemble de trois barres par une plaque et chaque ensemble de trois plaques par un grand cube (ce procédé ne permet pas de représenter des nombres à l'aide d'unités d'ordre supérieur au quatrième ordre) ;
- le « compteur humain » binaire (jeu présenté par T. L. Fletcher in *L'Apprentissage de la mathématique aujourd'hui)* : « Plusieurs enfants sont alignés (les mains baissées). Il leur est précisé que, dans la suite, leur main droite doit être nettement dirigée vers le haut ou vers le bas. L'enfant situé le plus à droite reçoit l'instruction de changer de position (du haut vers le bas ou du bas vers le haut) à chaque signal, un claquement de mains du professeur, par exemple ; les autres changent de position quand la main de l'enfant à leur gauche se dirige en bas. »

C'est là une réalisation pédagogique du principe même des compteurs.

Nombres à virgule

De nombreux codages utilisés dans la pratique sont fondés non pas sur l'ordre des entiers, mais sur celui des nombres à virgule : par exemple les cotes des livres dans les bibliothèques modernes, le numé-

rotage des maisons de certaines rues en *bis*, *ter*..

On y est amené lorsqu'on veut pouvoir intercaler des éléments entre deux éléments quelconques. Il s'agit d'un ordre analogue à celui des dictionnaires ; c'est pourquoi on l'appelle aussi « ordre lexicographique ». Cette question est en relation avec celle du repérage sur une demi-droite.

Construction de nombres à virgule binaires

Entre 0 et 1 on introduit un nombre noté « 0,1 » ; entre 0 et 0,1 on introduit un nombre noté « 0,01 » ; entre 0 et 0,01 on introduit un nombre noté « 0,001 » ; etc. De même, entre 1,1 et 1,11 on introduit un nombre noté « 1.101 », etc.

Un nombre à virgule binaire s'écrit donc comme un nombre entier en base deux suivi d'une virgule et d'une suite de « *0* » et de « 1 » en *nombre fini* avec la propriété que tout nombre est égal à tous ceux qu'on peut écrire en adjoignant des zéros à sa droite. Par exemple :

$$101{,}100 = 101{,}1000 = 101{,}1.$$

Nombres à virgule de base quelconque

On construit les nombres « décimaux » à partir des nombres naturels en introduisant neuf nouveaux nombres entre deux nombres naturels consécutifs, puis encore neuf nombres entre deux nombres consécutifs ainsi déterminés, et ainsi de suite.

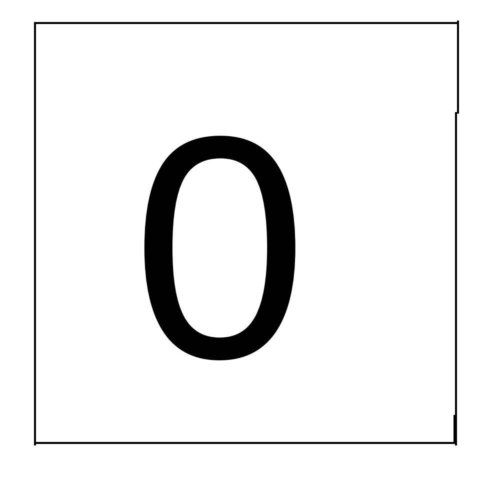
L'ensemble des entiers naturels apparaît donc comme sous-ensemble de l'ensemble des nombres décimaux, et ceux-ci sont écrits avec les symboles « 0 », « 1 », « 2 », « 3 », « 4 », « 5 », « 6 », « 7 », « 8 », « 9 » et le symbole de virgule « , ».

De même que les nombres à virgule binaires ou décimaux, on peut considérer des nombres à virgule de n'importe quelle base. Il faut cependant noter que, tandis que pour les entiers naturels changer de base revenait à changer le nom des mêmes nombres, ici la base intervient dans la définition des nombres eux-mêmes ; par exemple, le nombre à virgule ternaire 0,1 n'est pas un nombre décimal, car il s'agit du nombre rationnel 1/3 dont on sait que ce qu'on appelle le « développement décimal » s'écrit « 0,333...3... » avec une *infinité* de chiffres « 3 ».

JOSETTE ADDA

Bibliographie

J. CROSSLEY, *The Emergence of Number*, World Scientific Publ., River Edge (N. J.), 1987 / Z. P. DIENES, *Lu Mathématique moderne dans l'enseignement primaire*, O.C.D.L., Paris, 5' éd. 1970 / T. L. FLETCHER, *L'Apprentissage de la mathématique aujourd'hui. Une didactique nouvelle pour le second degré* (*Some Lessons in Mathematics*, 1965), trad. M. Glaymann et al.. O.C.D.L., 1966 / G. GUITEL, *Histoire comparée des numérations écrites*, Flammarion, Paris, 1975 / G. IFRAH, *Les Chiffres, ou l'Histoire d'une grande invention*, R. Laffont, Paris. 1985 / N. PICARD, *À lu conquête du nombre*, O.C.D.L., nouv. éd., 1975.



ORDONNÉS ENSEMBLES

Les relations d'ordre interviennent de manière naturelle dans des questions comme l'étude des liens de parenté et celle des liens de subordination, comme les problèmes de classification, etc. C'est de là, et de la relation $\leqslant$ entre nombres, que découle la terminologie habituellement employée : on dit que a est « plus petit » que b, que a est « dominé » par b, que b est « plus haut » que a, etc Remarquons que cette situation inclut l'égalité $a = b$; on précise que, de plus, $a \neq b$, en ajoutant l'adverbe « strictement ».

La théorie des ensembles ordonnés comporte une partie élémentaire qui est exposée ici, mais constitue aussi un chapitre important de la « grande » théorie des ensembles en liaison étroite avec l'axiome du choix dont plusieurs formulations équivalentes s'expriment en terme d'ordre. La théorie des ordinaux, due à Cantor, s'exprime aussi dans ce cadre.

Rappelons enfin que c'est à partir de la relation d'ordre usuel sur l'ensemble des nombres rationnels que R. Dedekind, en 1872, a donné la première construction rigoureuse de l'ensemble des nombres réels.



Relations d'ordre

On dit qu'une relation $\mathcal{R}$ sur un ensemble E est une relation d'ordre (cf. théorie élémentaire des ENSEMBLES, chap. 2) si elle satisfait aux axiomes suivants :

(O_1) Réflexivité : pour tout élément a de E, on a la relation $a\mathcal{R}a$;

(O_2) Antisymétrie : les relations $a\mathcal{R}b$ et $b\mathcal{R}a$ ne sont compatibles que pour $a = b$;

(O_3) Transitivité : les relations $a\mathcal{R}b$ et $b\mathcal{R}c$ impliquent $a\mathcal{R}c$.

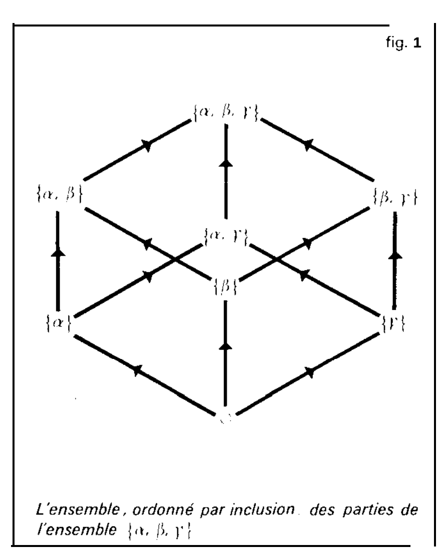
Par exemple, la relation $\leqslant$ est une relation d'ordre sur tout sous-ensemble de l'ensemble R des nombres réels.

Étant donné deux nombres réels distincts a et b, on a toujours une, et une seule, des relations $a \leqslant b$ *ou* $b \leqslant a$. Plus généralement, si E est un ensemble muni d'une relation d'ordre $\mathcal{R}$, on dira que deux éléments a et b sont *comparables* si on a au moins une des relations $a\mathcal{R}b$ *ou* $b\mathcal{R}a$. Si deux éléments quelconques d'un ensemble ordonné E sont toujours comparables, on dit que E est totalement ordonné, ou encore que l'ordre est *total,* ou linéaire. Dans le cas contraire, on parle d'ordre *partiel.*

Un exemple d'ordre partiel

Soit X un ensemble ; la relation d'inclusion $\subset$ est une relation d'ordre sur l'ensemble $E = \mathfrak{P}(X)$ des parties de X (cf. théorie élémentaire des ENSEMBLES, chap. 1). Sur la

figure 1, on a représenté le *diagramme* ***sagittal*** de cette relation dans le cas où



L'ensemble, ordonné par inclusion, des parties de l'ensemble {α, β, γ}

X = {α, β, y} est un ensemble à trois éléments : pour ***a, b*** ∈ ***:I'(X),*** on a :

$$a \subset b,$$

si ***a*** = *b* ou s'il existe une ou plusieurs flèches « consécutives » du diagramme partant de ***a*** pour aboutir à *b*. Les parties { β} et {α, y}, par exemple, ne sont pas comparables et l'ordre n'est donc pas total.

intervalles

Soit E un ensemble ordonné et ***a*** **et** *b* des éléments de E avec ***a*** plus petit que *b*. On appelle *intervalle* ***ouvert*** d'origine ***a*** et d'extrémité *b*, noté ***]a*** *b*[, l'ensemble des éléments *x* **de** E qui sont strictement plus grands que a et strictement plus petits que *b* (donc en particulier comparables à *a* et à *b*). Ainsi on a, dans l'exemple précédent illustré par le diagramme de la figure 1,

$$]\{\alpha\}, \{\alpha, \beta, \gamma,\} [= \{\{\alpha, \beta\}, \{\alpha, \gamma\}\}.$$

On définit de même, pour *a* plus petit que ***b, l'intervalle*** *fermé* ***[a, b]*** qui est l'ensemble des éléments *x* de E qui sont plus grands que *a* et plus petits que ***b :*** c'est l'intervalle ouvert]*a*, ***b[*** augmenté de ses deux extrémités.

Majorants

Soit E un ensemble ordonné et A un sous-ensemble de E. On dit qu'un élément *m* de E est un *majorant* de A si tout élément de A est plus petit que *m* (ce qui implique que tout élément de A est comparable à *m*) ; si A admet au moins un majorant, on dit que c'est un ensemble ***majoré.*** Dans le cas où, de plus, *m* appartient à A, on dit que A admet un plus grand élément ; il résulte de (O_2) que ce plus grand élément *m*, s'il existe, est unique.

Reprenons encore l'exemple ci-dessus, illustré par le diagramme de la figure 1, avec :

$$\mathbf{A} = \{\varnothing, \{\alpha\}, \{\alpha, \beta\}, \{\beta, \gamma\}\} ;$$

cet ensemble admet pour majorant *m* = {α, β, γ}, mais n'admet pas de plus grand élément. L'élément {β, y} de A n'est pas un plus grand élément (car il n'est pas comparable à {α, β}), mais possède la propriété plus faible qu'il n'existe pas d'élément de A qui soit strictement plus grand que lui : on dit que c'est un élément *maximal.*

Borne supérieure

Soit toujours E un ensemble ordonné et A un sous-ensemble de E que nous supposerons majoré. Tout élément de E plus grand qu'un majorant de A est a fortiori un majorant de A et il est donc intéressant de chercher des majorants le plus petits possible. On dit que A admet une borne supérieure si l'ensemble de ses majorants a un plus petit élément ; cet élément,

nécessairement unique, s'appelle la *borne supérieure* de A et on le note :

$$\sup A.$$

Bien entendu, on définirait de même la notion de borne inférieure.

Dans le cas des nombres réels, tout ensemble majoré admet une borne supérieure (cf. CALCUL INFINITÉSIMAL - Calcul à une variable, chap. 1) et c'est cette importante propriété qui permet de « faire de l'analyse », alors que l'existence de lacunes dans l'ensemble des nombres rationnels met cette propriété en défaut : c'est ainsi que l'ensemble des nombres rationnels dont le carré est strictement inférieur à 2 n'admet pas de borne supérieure dans Q ; dans R, sa borne supérieure est le nombre (irrationnel) $\sqrt{2}$.

Lorsque l'ordre est total, tout sous-ensemble fini a un plus grand élément ; mais il n'en est plus de même si l'ordre est partiel. Étant donné deux éléments a et *b*, on désigne par :

$$\sup(a, b)$$

la borne supérieure (si elle existe) de l'ensemble $\{a, b\}$. Par exemple, si l'ensemble E = $\mathcal{P}(X)$ des parties d'un ensemble X est ordonné par inclusion, soit a et *b* deux parties de E. Une partie c est « plus grande » que a et que b, c'est-à-dire $a \subset c$ et $b \subset c$, si, et seulement si, $a \cup b \subset c$; la réunion $a \cup b$ est donc le plus petit majorant commun à a et à *b*, c'est-à-dire la borne supérieure. Raisonnant de même pour la borne inférieure, on a donc dans notre exemple :

$$\sup(a, b) = a \cup b, \quad \inf(a, b) = a \cap b.$$

On appelle *treillis* tout ensemble ordonné dans lequel deux éléments quelconques ont toujours une borne supérieure et une borne inférieure ; la théorie des treillis est une branche de l'algèbre qui a de nombreuses applications tant en mathématiques pures qu'en mathématiques appliquées.

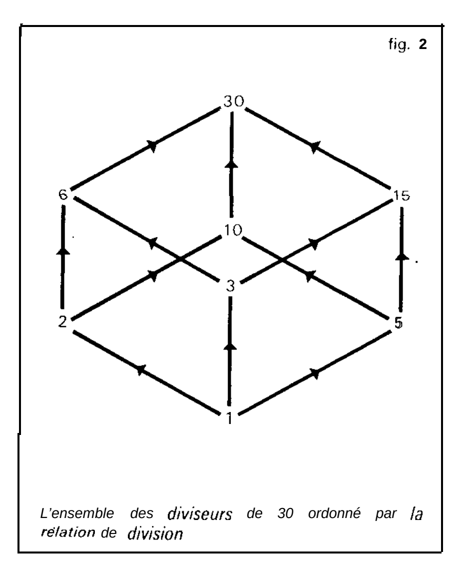
Quelques ordres sur N*

On peut munir l'ensemble **N*** des entiers naturels strictement positifs de diverses relations d'ordre qui montreront bien la grande variété de propriétés que l'on peut obtenir ainsi.

Après la relation $\leqslant$ usuelle, la relation d'ordre la plus courante est la relation de divisibilité :

$$p \mid q,$$

si p divise q, c'est-à-dire si q est multiple de p : cela signifie qu'il existe un entier $m \in$ N* tel que $q = mp$. Sur la figure 2, on



L'ensemble des diviseurs de 30 ordonné par la relation de division

a représenté le diagramme sagittal de l'ordre ainsi obtenu sur l'ensemble des huit diviseurs du nombre 30. On remarque la parfaite analogie avec l'ensemble des parties de X = $\{\alpha, \beta, \gamma\}$ ordonné par inclu-

sion ; plus généralement, on dit que deux ensembles ordonnés sont isomorphes s'il existe une bijection de l'un sur l'autre qui respecte les ordres.

La relation de divisibilité ne définit pas un ordre total, car 2 et 3, par exemple, ne sont pas comparables. Si p et q sont deux entiers. l'ensemble des « majorants » communs est l'ensemble des multiples communs dep et de q ; cet ensemble a un « plus petit » élément, le plus petit commun multiple (P. P. C. M.) de p et de q, qui est donc la borne supérieure de p et de y pour la relation de divisibilité. On interpréterait de même le P. G. C. D. (plus grand commun diviseur) de p et de y comme la borne inférieure de p et de q, ce qui donne une structure de treillis.

Très différente est la relation d'ordre suivante sur **N***, qui intervient dans une démonstration du théorème de d'Alembert sur les racines d'une équation algébrique. On remarque d'abord que tout entier $n \geqslant 1$ s'écrit de manière unique sous la forme :

$$n = 2' \ (26 + 1),$$

avec a, $b \in$ N*. Cela signifie que n est divisible par 2^a, mais pas par 2^{a+1} ; le quotient de n par 2^a est alors un entier impair, donc de la forme $2b + 1$. Par exemple, 72 est divisible par $8 = 2^3$ mais n'est pas divisible par $16 = 2^4$; le quotient de 72 par 8 est le nombre impair $9 = 4 \times 2 + 1$, d'où l'on a $72 = 2^3(4 \times 2 + 1)$.

Pour $n = 2^a (2b + 1)$ et $n' = 2^{a'}(2b' + 1)$ on dira que n est dominé par n', et on notera $n \leqslant n'$, si on a :

$$a < a'$$

ou si on a simultanément :

$$a = a', \quad b \leqslant b' ;$$

on définit ainsi une relation d'ordre total très différente de l'ordre usuel. Par exemple $31 \leqslant 2$ ou $12 \leqslant 8$. L'intervalle ouvert]2, 14[contient les nombres 6 et 10, mais l'intervalle]2, 12[contient une infinité d'éléments, à savoir **tous** les nombres de la forme $2(2k + 1)$, $k \geqslant 1$, et le nombre 4 qui est son plus grand élément.

L'ordre lexicographique

Un ordre important dans les applications les plus variées (pour tous les problèmes de classification en sciences humaines, par exemple) est l'ordre lexicographique. Il est familier à tous ceux qui ont consulté un dictionnaire.

Soit X un ensemble ordonné par $\leqslant$ que nous appellerons un alphabet. On appelle ***mot*** toute suite finie d'éléments de X, sans se préoccuper du sens éventuel de ce mot dans une langue naturelle. Par exemple, si X est l'alphabet usuel, constitué par nos vingt-six lettres,

abbcda, *encyclopedie*

sont des mots.

L'ordre lexicographique se définit alors sur l'ensemble E des mots de la manière suivante. Si $x = x_1x_2 \ x_p$ et $y = y_1y_2 \ y_q$ sont des mots, on dira que :

$$x \leqslant y$$

si on a $p \leqslant q$ et $x_1 = y_1$, $x_2 = y_2$, ... $x_p = y_p$, ou si, désignant par k le plus petit entier tel que $x_k \neq y_k$, on a $x_k \leqslant y_k$. Ainsi, si l'un des deux mots n'est pas obtenu en rajoutant des lettres à l'autre, on classe ces mots en examinant la première lettre qui diffère, par exemple :

aba $\leqslant$ ***abaa*** $\leqslant$ ***abd.***

ANDRÉ WARUSFEL

Bibliographie

M. BARBUT & B. MONJARDET, Ordre *et classification*, 2 vol., Paris, 1971 / A. BOUVIER, *La Théorie des ensembles*, coll. Que sais-je ?, P.U.F., Paris, 3' éd. 1982 / N. BOURBAKI, *Théorie des ensembles*, Masson, Paris, nouv. éd. 1990 / P. R. HALMOS, *Introduction à lu théorie des ensembles,* Paris, 1967 / J. PICHON, *Théorie des ensembles, logique, les entiers,* éd. Marketing, Paris, 1989.

ORTHOGONAUX POLYNÔMES

C'est à travers l'étude de certains problèmes d'analyse fonctionnelle (équations intégrales, skies de Fourier, problème de Sturm-Liouville et, plus généralement, problèmes aux limites dans les équations aux dérivées partielles) qu'est apparue la notion de système orthogonal de fonctions. Ces problèmes amènent à considérer des espaces hermitiens constitués de fonctions et à déterminer les valeurs propres et les fonctions propres (cf. théorie SPECTRALE) de certains endomorphismes de ces espaces. Dans le cas d'un opérateur hermitien, les sous-espaces propres sont orthogonaux deux à deux. Le problème essentiel consiste alors à chercher des bases hilbertiennes constituées de fonctions propres.



Équation intégrale de Fredholm

Soit E un ensemble muni d'une mesure positive μ et k une fonction de carré intégrable sur E X E. Pour toute fonction f de carré intégrable sur E et pour presque tout élément x de E, la fonction $y \mapsto k(x, y) f(y)$ est intégrable sur E et la fonction g, définie presque partout par la formule :

$$g(x) = \int_E k(x,y) f(y)\, d\mu(y),$$

est de carré intégrable sur E. L'application U_k, dite associée au noyau k, qui à tout élément f de L'(E) associe g, est un endomorphisme de L'(E). Lorsqu'on munit $L^2(E)$ de la norme de la convergence en moyenne quadratique, cet endomorphisme est continu et sa norme est inférieure à $\|k\|_2$; cet endomorphisme est même un endomorphisme compact. La résolution de l'équation intégrale de Fredholm :

$$(1) \quad \int_E k(x,y) f(y)\, d\mu(y) - \lambda f(x) = h(x),$$

où h est un élément donné de $L^2(E)$, conduit à chercher les valeurs propres et les vecteurs propres de I'endomorphisme U_k. Lorsque le noyau k est hermitien, c'est-à-dire lorsque, pour **tout** couple (x, y) d'éléments de E, $k(y, x) = \overline{k(x, y)}$, alors l'endomorphisme compact U_k est hermitien. La théorie spectrale montre que l'ensemble $sp(U_k)$ des valeurs propres de U_k est une partie bornée dénombrable de R, dont **tous** les points, sauf peut-être 0, sont isolés. De plus, les sous-espaces propres E_λ sont orthogonaux deux à deux et le sous-espace vectoriel :

$$F = \bigoplus_{\lambda \in sp(U_k)} E_\lambda$$

est dense dans $L^2(E)$. Enfin, E est de dimension finie si A $\neq$ 0. Il existe donc une suite (A,,) de nombres réels convergeant vers 0 et une base hilbertienne (φ_n) de $L^2(E)$ telles que, pour tout entier n, $U_k(\varphi_n) = \lambda_n \varphi_n$. Une telle base (φ_n)

s'appelle système orthogonal associé au noyau k. Enfin, la suite (λ_n) est de carré sommable :

$$\sum_{n=0}^{+\infty} \lambda_n^2 \leqslant \|k\|_2^2,$$

et le noyau k peut se développer de la manière suivante :

$$k(x, y) = \sum_{n=0}^{+\infty} \lambda_n \varphi_n(x) \overline{\varphi_n(y)}.$$

Pour résoudre l'équation intégrale (1), on décompose le second membre h dans la base hilbertienne précédente :

$$h = \sum_{n=0}^{+\infty} \alpha_n \varphi_n, \qquad \alpha_n = (h \mid \varphi_n).$$

Pour que :

$$f = \sum_{n=0}^{+\infty} \beta_n \varphi_n$$

soit solution de (1), il faut et il suffit que, pour tout entier naturel n,

$$(\lambda_n - \lambda)\beta_n = \alpha_n.$$

En particulier, lorsque λ n'appartient pas à $\mathrm{sp}(U_k) \cup \{0\}$, l'équation (1) admet une solution et une seule. Lorsque $\lambda \in \mathrm{sp}(U_k) \cup \{0\}$, pour que (1) admette une solution, il faut et il suffit que h soit orthogonale au sous-espace vectoriel E_λ. Enfin, lorsque $\lambda = 0$, pour que (1) admette une solution, il faut et il suffit que h soit orthogonale au noyau de U_k et que :

$$\sum_{n \in P} \frac{|\alpha_n|^2}{\lambda_n^2} < +\infty,$$

où P désigne l'ensemble des entiers n tels que $\lambda_n \neq 0$.

On notera que les séries précédentes convergent en moyenne quadratique. E. Schmidt (1907) et T. Mercer (1909) ont trouvé des conditions assez larges sous lesquelles la convergence est uniforme.

La théorie spectrale d'opérateurs hermitiens plus généraux conduit encore à des théories analogues. Signalons le cas des séries de Fourier (cf. analyse HARMONIQUE, espace de HILBERT) et celui des fonctions sphériques (cf. GROUPES - Groupes de Lie). Nous allons nous borner ici à un cas particulièrement simple.

Polynômes orthogonaux

Soit I un intervalle de R non réduit à un point et p une fonction à valeurs réelles continue sur I, telle qu'en tout point x intérieur à I, $p(x) > 0$. Soit $\mathcal{C}_I(p)$ l'espace vectoriel des fonctions f à valeurs complexes continues sur I telles que :

$$\int_I |f(x)|^2 p(x)\, dx < +\infty.$$

On munit $\mathcal{C}_I(p)$ du produit hermitien :

$$(f, g) \mapsto (f \mid g) = \int_I f(x)\overline{g(x)}p(x)\, dx.$$

L'espace hermitien $\mathcal{C}_I(p)$ n'étant pas complet, on est amené à le considérer comme un sous-espace vectoriel $L_I^2(p)$ des classes de fonctions f mesurables sur I à valeurs complexes et telles que :

$$\int_I |f(x)|^2 p(x)\, dx < +\infty.$$

Muni du produit hermitien précédent, $L_I^2(p)$ est un espace hilbertien.

Plaçons-nous dans l'un des deux cas suivants :

a) L'intervalle I est borné et p est intégrable sur I, c'est-à-dire que :

$$\int_I p(x)\, dx < +\infty.$$

b) L'intervalle 1 est non borné, p est intégrable sur I et à décroissance rapide à l'infini, c'est-à-dire que, pour tout entier n,

$$\lim_{x \to \pm\infty} x^n p(x) = 0$$

Les fonctions monomiales $e_n : x \mapsto x^n$ appartiennent alors à $\mathcal{C}_I(p)$. La suite (P,) des fonctions polynomiales déduite de la famille (e,) par orthonormalisation est appelée *système* ***depolynômes orthogonaux*** sur I associé au poids ***p*** ; pour tout entier naturel n, la suite (P,) est un polynôme à coefficients réels de degré n, et le coefficient dominant de (P,) est strictement positif.

Réciproquement, soit (Q,) une suite orthogonale de polynômes à coefficients complexes telle que, pour tout entier n, le polynôme Q_n soit de degré n. Pour tout entier n, il existe un nombre complexe λ_n et un seul tel que $Q_n = \lambda_n P_n$; plus précisément :

$$\lambda_n = (Qn \; P_n).$$

En utilisant le fait que P_n est orthogonal à tout polynôme de degré inférieur ou égal à $n \quad 1$, on prouve facilement les résultats suivants :

Pour tout entier naturel non nul n, il existe un triplet (α_n, β_n, y,,) de nombres réels et un seul tel que :

$$P_{n+1}(x) = (\alpha_n x + \beta_n) P_n(x) + \gamma_n P_{n-1}(x)$$

(formule de récurrence linéaire à deux termes) ; en outre, α_n est strictement positif et γ_n strictement négatif.

Toutes les racines de P_n sont réelles, simples et intérieures à 1 et, pour tout entier naturel non nul n, les racines de P_n séparent celles de P_{n+1}.

Enfin, lorsque l'intervalle I est symétrique par rapport à 0 et que la fonction p est paire, le polynôme P,, est pair si n est pair, impair si n est impair, et $\beta_n = 0$.

Il reste à examiner si la suite (P,,) est une base hilbertienne ou, ce qui revient au même. si le sous-espace vectoriel engendré par les fonctions e,, est dense dans C,(p).

Lorsque l'intervalle I est borné, il en est toujours ainsi ; cela résulte du théorème d'approximation de Weierstrass et du fait que, sur un intervalle borné, la convergence uniforme implique la convergence dans $\mathcal{C}_I(p)$.

Lorsque l'intervalle I n'est pas borné, il peut arriver que (P,) ne soit pas une base hilbertienne, par exemple si $p(x) =$ exp (- $|x|^\alpha$), où $\alpha \in]0, 1[$. Cependant. lorsque ***p*** est à décroissance exponentielle. c'est-à-dire lorsque ***p*** est dominée par une fonction de la forme $x \mapsto$ exp $(-\alpha \quad x \quad)$, où $\alpha > 0$, au voisinage de $\pm\infty$, la suite est une base hilbertienne de $L^2_I(p)$ et a fortiori de C,(p). En effet, tous les moments $M_n = (f \quad e_n)$ d'un élément f de $L^2_I(p)$ orthogonal aux polynômes P,, sont nuls. D'autre part, la décroissance exponentielie du poids ***p*** permet de prouver que la bande de convergence de la transformée de Laplace de fp est non vide. On en déduit alors que fp est nulle presque partout et que f est nulle presque partout. Le problème de la recherche de conditions portant sur ***p*** pour que la suite (P,,) soit une base hilbertienne (problème de Bernstein) est assez délicat ; il a fait l'objet de travaux de A. Denjoy (1922) et de T. Carleman (1932) et, plus récemment, de W. Pollard (1956) et de J.-P. Ferrier (1965), qui ont obtenu des conditions nécessaires et suffisantes.

Équations différentielles des polynômes orthogonaux

Soit $1 = [a, \beta]$ un intervalle compact de R, ***a*** et *b* deux fonctions à valeurs réelles indéfiniment dérivables sur I, la fonction a ne s'annulant pas sur l'intérieur de I et

admettant un zéro simple aux points α et β. On considère l'équation différentielle :

$$(1) \qquad ay'' + by' = \lambda y,$$

où A est un nombre complexe. De telles équations interviennent, par exemple, dans les problèmes de Sturm-Liouville. Les solutions de (1) sont les fonctions propres de l'endomorphisme U $: f \mapsto af'' + bf'$ de l'espace vectoriel E des fonctions indéfiniment dfrivablcs sur I. Pour étudier l'équation (1), on introduit sa fonction résolvante, c'est-à-dire une fonction Y à valeurs réelles strictement positives, définie sur l'intérieur de I vérifiant l'équation différentielle :

$$(ru)' = rb\ ;$$

alors :

$$U(f) = \frac{1}{r}(raf')'.$$

Supposons que les nombres :

$$\mu = \frac{b(\alpha)}{a'(\alpha)}, \quad \upsilon = \frac{b(\beta)}{a'(\beta)}$$

soient réels strictement positifs. Dans ce cas, $(x - \alpha)^{\mu} r(x)a(x)$ admet une limite finie non nulle au point α et $(\beta - x)^{\upsilon} r(x)a(x)$ admet une limite finie non nulle au point β. Par suite, pour tout couple (f, g) d'éléments de E, la fonction $rf\bar{g}$ est intégrable sur I.

On peut donc définir un produit hermitien sur E par la formule :

$$(f \mid g) = \int_{\alpha}^{\beta} r(x)f(x)\overline{g'(x)}dx.$$

L'endomorphisme U est alors hermitien ; plus précisément :

$$(U(f) \mid g) = (f \mid U(g))$$
$$= -\int_{\alpha}^{\beta} r(x)a(x)f'(x)g(x)dx.$$

Dans beaucoup de cas intervenant en pratique, on peut déterminer une base hilbertienne de E constituée de vecteurs propres de U. Nous nous contentons ici d'examiner le cas où a et b sont des fonctions polynomiales de la forme suivante :

$$a(x) = (x - \alpha)(x - \beta),$$
$$b(x) = \gamma x + 6, \quad y \neq 0.$$

Pour tout entier naturel n, le sous-espace vectoriel E_n de E constitué des fonctions polynomiales de degré inférieur ou égal à n est stable par U. Les conditions $\mu > 0$ et $v > 0$ sont équivalentes aux conditions $\alpha\gamma + \delta > 0$ et $\beta\gamma + \delta > 0$. De plus :

$$r(x) = (x - \alpha)^{\mu-1}(\beta - x)^{\nu-1}$$

est une résolvante de U. Le système (P,) de polynômes orthogonaux associé au poids r est une base hilbertienne de E constituée de fonctions propres de U ; plus précisément :

$$U(P_n) = n(n + y - 1)P_n.$$

Les polynômes P,, s'appellent polynômes de Jacobi. Dans le cas où $\mu = v = 1$, on trouve les polynômes de Legendre ; dans le cas où $\mu = v = 1/2$, on trouve les polynômes de Tchebichev, ainsi que dans le cas où $\mu = v = 3/2$.

Soit maintenant I un intervalle de la forme $[\alpha, +\infty[$. On suppose cette fois que les fonctions a et b, ainsi que toutes leurs dérivées, sont des éléments de l'espace vectoriel E des fonctions à croissance lente au voisinage de $+\infty$, et on considère U comme un endomorphisme de E. On suppose que $\mu = b(\alpha) / a'(\alpha) > 0$ et que, d'autre part, $b(x) / a(x)$ admet une limite strictement négative v, finie ou infinie, lorsque x tend vers $+\infty$. Pour tout couple (f, g) d'éléments de E, la fonction $rf\bar{g}$ est

alors intégrable sur I, et U est encore hermitien pour le produit hermitien précédemment défini. Lorsque $a(x) = x$ a et que h(s) = $\gamma x + \delta$, avec y $\neq 0$, les conditions $\mu > 0$ et $\nu < 0$ sont équivalentes aux conditions $\gamma < 0$ et $\gamma\alpha + \delta > 0$. De plus,

$$r(x) = (x - \alpha)^{\mu - 1} e^{\gamma x}$$

est une résolvante de x. Le système (P_n) de polynômes orthogonaux associé au poids r est une base hilbertienne de E constituée de fonctions propres de U ; plus précisément :

$$U(P_n) = n\gamma P_n ;$$

les polynômes P_n s'appellent polynômes de Sonine ; dans le cas où $\mu = 1$, on trouve les polynômes de Laguerre.

Examinons enfin le cas où $I = R$; on suppose que les fonctions a et b, ainsi que toutes leurs dérivées, sont des éléments de l'espace vectoriel E des fonctions à croissance lente au voisinage de $\pm\infty$, et on considère U comme un endomorphisme de E ; on suppose de plus que $b(x) / a(x)$ admet des limites $\mu > 0$ et $\nu < 0$, finies ou infinies, lorsque x tend vers $-\infty$ et vers $+\infty$. La théorie se poursuit alors comme dans les cas précédents. Lorsqu'on a :

$$a(x) = 1, \quad b(x) = \gamma x + \delta, \quad \delta \neq 0,$$

les conditions $\mu > 0$ et $\nu < 0$ sont équivalentes à la condition y < 0. De plus :

$$r : x \mapsto \exp\left(\frac{\gamma}{2} x^2\right)$$

est une résolvante. Le système (P_n) des polynômes orthogonaux associé au poids r est encore une base hilbertienne de E constituée de fonctions propres de U ; plus précisément :

$$U(P_n) = \gamma n P_n ;$$

les polynômes P_n s'appellent alors polynômes d'Hermite.

Fonctions génératrices des polynômes orthogonaux

Les polynômes orthogonaux P_n précédemment introduits peuvent se calculer de la manière suivante : de la relation $(ru)' = rb$, on déduit, par récurrence sur n, que :

$$D^n(ra^n) = rQ_n,$$

où Q_n est une fonction polynomiale de degré n. Par intégrations par parties, on prouve que, pour tout entier n, Q_n est proportionnel à P_n : c'est la formule de Rodrigues. De plus, la résolvante r peut se prolonger en une fonction holomorphe sur C {a, β}. La formule intégrale de Cauchy permet alors d'établir la formule de Schläffli :

$$r(z)Q_n(z) = \frac{n!}{2i} \int_\Gamma \frac{r(\zeta)a(\zeta)^n}{(\zeta - z)^{n+1}} d\zeta,$$

où $z \in C - \{\alpha, \beta\}$ et où Γ est un cercle d'indice 0 par rapport à a et β. On en déduit le résultat suivant (fonction génératrice des polynômes orthogonaux).

Soit x un point de I, et ρ un nombre réel strictement positif tel que le cercle Γ de centre x et de rayon ρ soit d'indice 0 par rapport à a et β. Pour tout nombre complexe u tel que $u \sup_{\zeta\in\Gamma} a(\zeta) < p$,

$$\sum_{n=0}^{+\infty} \frac{Q_n(x)}{n!} u^n = \frac{r(w)}{r(x)} \frac{1}{1 - ua'(w)},$$

où w est le seul élément du disque ouvert de centre x et de rayon ρ tel que :

$$w - x - ua(w) = 0.$$

Dans le cas des polynômes de Legendre réduits, c'est-à-dire le cas où $a(x) = x^2 - 1$ et où h(x) = 2 x, on peut prendre $r = 1$; le polynôme Q,, satisfait alors à l'équation différentielle :

$$(x^2 - 1)y'' + 2\,xy' \quad n(n+1)y = 0 ;$$

d'où :

$$Q_n(x) = D^n[(x^2 - 1)^n],$$

et :

$$\sum_{n=0}^{+\infty} \frac{Q_n(x)}{n!} u^n = \frac{1}{\sqrt{4u^2 - 4xu + 1}};$$

lorsque $|u| < 1/6$, cette strie converge uniformément sur [- 1, 1].

De même, la fonction génératrice des polynômes de Laguerre réduits, c'est-à-dire dans le cas où $a(x) = x$ et où $b(x) = 1 - x$, est :

$$\sum_{n=0}^{+\infty} \frac{Q_n(x)}{n!} u^n = \frac{1}{1-u} \exp\left(-\frac{ux}{1-u}\right).$$

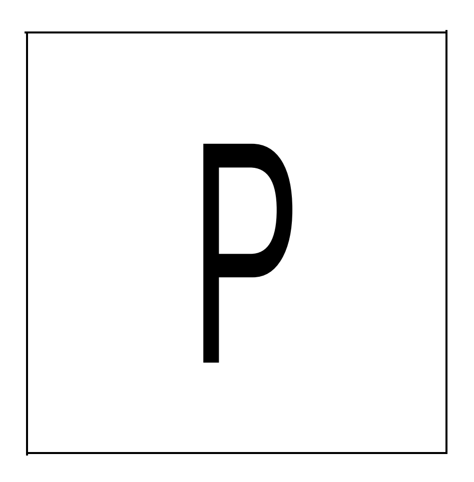
Enfin, la fonction génératrice des polynômes d'Hermite réduits, c'est-à-dire dans le cas où $a(x) = 1$ et $b(x) = 2x$, est :

$$\sum_{n=0}^{+\infty} \frac{Q_n(x)}{n!} u^n = = \exp(-u^2 + 2ux).$$

JEAN-LOUIS OVAERT

Bibliographie

C. BREZINSKI. A. DRAUX, A. P. MAGNUS et al., *Polynômes orthogonaux et applications*, Springer, New York, 1985 / T. S. CHIHARA, *An Indroduction to Orthogonal Polynomials*, Cordon & Breach, New York, 1978 / J. DIEUDONNÉ, *Éléments d'analyse*, t. I, II et VI, Gauthier-Villars. Paris, 1962-1982 / S. GODOUNOV, *Équations de lu physique mathématique*, M.I.R., Moscou, 1973 / A. NIKIFOROV & V. UVAROV, *Special Functions of Mathematical Physics*, Birhauser Boston, Cambridge (Mass.), 1987 / V. SMIRNOV, *Cours de mathématiques supérieures*, t. II et III, *ibid.*, 1972-1982 / G. SZEGO, *Orthogonal Polynomials*, American Mathematical Society. Providence (R.I.), 1985.



P-ADIQUES NOMBRES

→ NOMBRES (THEORIE DES) - Nombres p-adiques

POLYNÔMES

La théorie des équations et des polynômes a i-té le propos essentiel de l'algèbre jusqu'au XIX[e] siècle (cf. ÉQUATIONS ALGÉBRIQUES, ALGÈBRE) et est à la base de la théorie des corps et de la théorie des nombres algébriques. Nous nous sommes limités ici à une construction formelle des objets mathématiques considérés, qui fait apparaître, sous le vocable « polynômes », l'existence de deux notions distinctes : les polynômes formels et les fonctions polynomiales. Cet article élémentaire pourra aussi servir d'introduction au maniement des notations abstraites.



Polynômes formels

La notion de polynôme est familière, mais on s'est contenté pendant fort longtemps de décrire des règles de calcul sans définir véritablement les objets mathématiques considérés. On trouve couramment des définitions comme : « Un monôme entier en la variable x est une expression de la forme Ax^n, A étant un coefficient numérique et n un entier positif » ; « Un polynôme en la variable x est une somme qui ne peut être composée (sic) que de nombres et de monômes entiers ». Puis suit l'énumération des règles de calcul sur ces objets.

La construction des polynômes donnée ici illustre, dans le cadre simple de l'algèbre élémentaire, la manière dont le mathématicien formalise, en suivant une voie qui peut sembler a priori déroutante, voire artificielle, certaines notions tenues pour « évidentes » ou « intuitives ».

Définition

Soit A un anneau commutatif unitaire (cf. ANNEAUX ET ALGÈBRES). On appelle *polynôme* à une *indéterminée* (cette terminologie sera justifiée plus loin) à *coefficients dans* A toute suite :

$$P = (a_0, a_1, a_2, \ldots, a_n, \ldots)$$

d'élément de A nuls sauf au plus un nombre fini d'entre eux (c'est-à-dire tous nuls à partir d'un certain rang). Les éléments a_i sont les *coefficients* du polynôme P.

Les polynômes étant définis comme des cas particuliers de suites (c'est-à-dire d'applications de l'ensemble N des entiers naturels dans A), deux tels polynômes sont donc égaux si et seulement s'ils ont les mêmes coefficients.

Anneau des polynômes

Nous allons maintenant définir formellement l'addition et la multiplication. Soit :

$$P = (a_0, a_1, \ldots, a_n, \ldots),$$
$$Q = (b_0, b_1, \ldots, b_n, \ldots),$$

deux polynômes à coefficients dans A. On appelle somme et produit de P et Q respectivement les polynômes :

$$(1) \quad P + Q = (a_0 + b_0, a_1 + b_1, \ldots, a_n + b_n, \ldots)$$

et :

$$(2) \quad PQ = (c_0, c_1, \ldots, c_n, \ldots).$$

avec :

$$c_0 = a_0b_0, c_1 = a_0b_1 + a_1b_0, \ldots, c_n = a_0b_n + a_1b_{n-1} + \ldots + a_nb_0.$$

Il est facile de vérifier que l'ensemble des polynômes considérés est ainsi muni d'une structure d'anneau commutatif unitaire ; nous désignerons provisoirement cet anneau par L. On va montrer qu'on peut « identifier » A à un sous-anneau de l'anneau L en remarquant pour cela que l'application :

$$a \mapsto (a, 0, 0, \ldots, 0, \ldots)$$

est un *isomorphisme d'anneau* de A sur le sous-anneau A' de L formé des polynômes dont tous les coefficients de rang $\geqslant 2$ sont nuls. Il est donc équivalent de « calculer » dans A ou de faire ces calculs sur les éléments correspondants de A', et nous identifierons ces deux anneaux en utilisant l'écriture abrégée :

$$a = (a, 0, \ldots, 0, \ldots).$$

pour tout $a \in$ A. Pour cette raison, les éléments de A' sont appelés des *constantes, ou* des *polynômes constants.*

Remarquons que, si a est un polynôme constant et P un polynôme quelconque, la

multiplication de P par a revient à multiplier tous les coefficients de P par a. Dans le cas où A est un corps, cette « multiplication scalaire » $(a, P) \mapsto aP$ munit l'anneau des polynômes à coefficients dans A d'une structure d'algèbre commutative sur le corps A.

Notion d'indéterminée

Désignons par X le polynôme :

$$x = (0, 1, 0, \ldots . 0, \ldots),$$

dont tous les coefficients sont nuls, sauf le second coefficient, qui est égal à l'élément 1 de l'anneau A. Il résulte de la définition (2) de la multiplication que l'on a :

$$X^2 = (0, 0, 1, 0, \ldots . 0, \ldots).$$
$$X^3 = (0, 0, 0, 1, 0, \ldots . 0, \ldots)$$

et, plus généralement, pour tout entier $n > 0$,

$$X^n = (\delta_{0n}, \delta_{1n}, \ldots, \delta_{nn}, \ldots).$$

où δ_{ij} est le symbole de Kronecker (égal à 1 si $i = j$ et à 0 si $i \neq j$). Si on a:

$$P = (a_0, a_1, \ldots, a_n),$$

avec $a_p = 0$ pour $p > n$ et $a_n \neq 0$, on obtient donc, en appliquant les définitions :

$$(3) \quad P = a_0 + a_1X + a_2X^2 + + a_nX^n = \sum_{i=0}^{n} a_iX^i, \; a_n \neq 0.$$

Le nombre n qui figure dans la formule (3) s'appelle le *degré* du polynôme P, noté d"(P) ; tout polynôme différent du polynôme nul a un degré bien déterminé, et l'écriture (3), appelée développement de P suivant les puissances croissantes, est unique. L'anneau des polynômes étant commutatif, on pourrait tout aussi bien « ordonner P suivant les puissances décroissantes » en écrivant :

$$P = a_nX^n + + a_1X + a_0, \quad a_n \neq 0.$$

Pour éviter des cas d'exception, on pose souvent $d^0(0) = \infty$, symbole formel régi par les conventions suivantes : pour tout entier naturel n, on pose $-\infty < n$ et $n + (-\infty) = -\infty$; $(-\infty) + (-\infty) = -\infty$. On peut alors énoncer, quels que soient les polynômes P et Q, des résultats tels que :

$$d^0(P + Q) \leqslant \sup(d^0(P), d^0(Q)),$$

avec égalité si d"(P) $\neq d^0(Q)$, et :

$$d^0(PQ) \leqslant d^0(P) + d^0(Q),$$

avec égalité si A est un anneau intègre.

La notation (3) justifie la terminologie de polynôme « à une indéterminée » et la notation A[X] que l'on utilise pour désigner l'anneau des polynômes à une indéterminée à coefficients dans A. Il est clair que la lettre X que l'on utilise pour désigner le polynôme (0, 1, 0, . ..) est arbitraire, en ce sens que, si, dans un texte mathématique, les lettres X et Y sont « disponibles », c'est-à-dire si elles n'ont pas encore été employées précédemment, on a A[X] = A[Y].

La construction formelle est maintenant terminée et les polynômes sont complètement définis. Il nous suffit de retenir que ce sont des objets mathématiques qui s'écrivent de manière unique sous la forme (3) et qui obéissent aux règles usuelles de calcul dans un anneau.

Remarquons pour terminer que, si A est un corps,

$$1, x, X^2, \ldots . X^n$$

est une base de l'espace vectoriel (sur A) des polynômes de dimension $\leqslant n$, espace qui est donc de dimension $n + 1$.

Dérivation formelle

L'examen des règles classiques de dérivation des fonctions numériques conduit à une approche formelle de la dérivation

dans un anneau. On appelle dérivation d'un anneau commutatif unitaire B une application :

$$D : B - B$$

telle que, quels que soient P et Q $\in$ B, on ait :

(4) $D(P + Q) = D(P) + D(Q)$,
(5) $D(PQ) = D(P)Q + PD(Q)$.

On s'intéressera ici à l'unique dérivation de B = A[X] telle que :

$$D(X) = 1, \; D(a) = 0,$$

pour tout polynôme constant $a \in$ A. Si P = a, + a, X + ... + $a_n X^n$, il résulte immédiatement des conditions (4) et (5) que l'on a :

$$D(P) = a, + 2a_2 X + ___ + na_n X^{n-1};$$

on note souvent D(P) = P'. Par récurrence, on alors la définition :

$$P'' = D^2(P), \ldots . P^{(k)} = D^k(P)$$

Polynômes à plusieurs indéterminées

Si A est un anneau commutatif unitaire, il en est de même de l'anneau B = A[X] des polynômes à une indéterminée à coefficients dans A. On peut donc considérer l'anneau C des polynômes à une indéterminée à coefficients dans B, soit :

$$C = B[Y] = (A[X])[Y];$$

il faut employer une autre lettre, Y, car X a déjà été utilisé. L'anneau B s'identifie à un sous-anneau de C et tout élément non nul de C s'écrit de manière unique :

$$\sum_{j=0}^{q} P_j Y^j, \; P_q \neq 0,$$

où P, $\in$ A[X], soit :

$$P_j = \sum_{i=0}^{p} a_{ij} X^i ;$$

d'après les règles de calcul dans l'anneau C, tout polynôme non nul de C s'écrit donc de manière unique sous la forme :

$$(6) \qquad \sum_{i=0}^{p} \sum_{j=0}^{q} a_{ij} X^i Y^j, a_{pq} \neq 0$$

La symétrie qui apparaît dans cette formule suggère la notation symétrique C = A[X,Y] pour désigner l'anneau des polynômes à deux indéterminées à coefficients dans A. On définit de la manière usuelle le degré total d'un tel polynôme, égal à $p + q$ si le polynôme est sous la forme (6), les degrés partiels en X et en Y, égaux respectivement à p et y dans (6), et les dérivations partielles formelles.

La construction précédente s'étend sans difficulté par récurrence pour définir l'anneau $A[X_1, \ldots . X_n]$ des polynômes à n variables à coefficients dans A.

Si A est un anneau d'intégrité, les anneaux de polynômes $A[X_1, \ldots . X_n]$ sont des anneaux d'intégrité. Dans le cas particulier où A = K est un corps commutatif, le corps des fractions (cf. ANNEAUX COMMUTATIFS, chap. 1, et CORPS, chap. 2) de l'anneau $K[X_1, \ldots . X_n]$ est le corps des fractions rationnelles à n variables à coefficients dans K ; on le note traditionnellement $K(X_1, \ldots . X_n)$.

Division euclidienne

Nous supposerons dans ce qui suit que A = K est un corps commutatif. L'anneau K[X] possède alors des propriétés arithmétiques très voisines de celles de l'anneau Z des entiers relatifs. Cela traduit le fait que l'un et l'autre sont des anneaux principaux et on peut dire que cette notion unificatrice d'anneau principal est née essentiellement de la répétition parfaite, pour l'anneau K[X], de toutes les considérations de divisibilité valables dans **Z**

(Cf. ANNEAUX COMMUTATIFS, chap. 2). Comme pour Z, la démonstration du fait que tout idéal est principal repose sur l'existence d'une *division euclidienne :* Si A et B $\in$ K[X], il existe des polynômes Q et R déterminés de manière unique tels que :

$$(7) \qquad A = BQ + R, \quad d^0(R) < d^0(B) \; ;$$

le cas R = 0 exprime que A est un multiple de B. On voit ici l'intérêt de la convention $d^0(0) = -\infty$, qui nous évite un cas d'exception.

Soit I un idéal de K[X]. Nous pouvons supposer $I \neq \{0\}$ et nous choisissons dans I un polynôme B $\neq$ 0 de degré minimum. Soit A $\in$ I ; la division euclidienne (7) de A par B entraîne que R $=$ A $-$ BQ appartient à I puisque I est un idéal qui contient A et B. L'inégalité $d^0(R) < d^0(B)$ entraîne R = 0 puisque B est de degré minimum parmi les polynômes $\neq$ 0. Ainsi l'idéal I est formé des multiples du polynôme B et est donc principal.

La structure arithmétique des anneaux de polynômes à plusieurs indéterminées est plus compliquée et nous renvoyons à l'article ANNEAUX COMMUTATIFS pour des indications sur ce sujet.

Fonctions polynomiales

À l'exception de tout ce qui concerne les racines, les résultats qui seront énoncés dans le présent chapitre s'étendent facilement au cas des polynômes à plusieurs indéterminées ; nous nous contenterons de les énoncer pour les polynômes à une indéterminée.

Fonction polynomiale associée à un polynôme formel

Soit A un anneau commutatif unitaire et :

$$P = \sum_{i=0}^{n} a_i X^i$$

un élément de A[X] écrit sous la forme (3). On appelle valeur de P sur un élément $x \in$ A l'élément :

$$(8) \qquad P(x) = \sum_{i=0}^{n} a_i x^i \in A,$$

et *fonction polynomiale* associée à P l'application $P^* : A \longrightarrow A$ définie par $P^*(x) = P(s)$; dans la pratique, on désigne encore par P cette fonction polynomiale.

Les fonctions polynomiales, c'est-à-dire les applications de A dans A pouvant s'obtenir à partir des éléments de A[X], forment un anneau commutatif unitaire, et l'application de K[X] dans cet anneau qui à tout polynôme formel associe la fonction polynomiale correspondante est un homomorphisme (par définition surjectif) d'anneaux. Si A est un anneau d'intégrité infini, cet homomorphisme est en fait un *isomorphisme*, c'est-à-dire que deux polynômes P et Q $\in$ A[X] sont égaux si et seulement si $P(x) = Q(x)$ pour tout $x \in$ A. Pour obtenir un contre-exemple, il suffit de prendre pour A le corps fini {0, 1, 2} des classes d'entiers modulo 3 ; le polynôme non nul :

$$2X - 3X^2 + X^3 = X(X-1)(X-2)$$

prend la valeur 0 en tout point de A.

Remarque

Soit L un sur-anneau de l'anneau A. La formule (8) permet de définir P(x) pour tout x C L et de définir ainsi une application polynomiale. dite encore associée à P, de L dans L. Cette remarque va nous permettre de préciser un point de notation. Prenons pour L le sur-anneau A[X] , si Q $\in$ A[X], la notation P(Q) désigne un élément de A[X] qui s'obtient en « substi-

tuant Q à X » et en développant les puissances de Q obtenues, en tenant compte des règles de calcul dans A[X]. Si on prend, en particulier, Q = X, on obtient le polynôme P lui-même, soit P(X) = P, ce qui nous permet d'utiliser indifféremment, pour désigner un polynôme, la notation P ou la notation P(X).

Prenant l'anneau A[X, Y] des polynômes à deux indéterminées pour sur-anneau, on peut donc définir P(X + Y) $\in$ A[X, Y] pour tout P $\in$ K[X]. La « formule de Taylor » s'écrit ici, si P est de degré n,

$$P(X + Y) = P(X) + P'(X)Y + \frac{1}{2!}P''(X)Y^2 + \dots + \frac{1}{n!}P^{(n)}(X)Y^n$$

où les dérivations qui figurent sont les dérivées formelles définies au chapitre 1.

Racines

Soit P $\in$ A[X] et a un élément de A. La division euclidienne de P par X-a s'écrit :

$$P = (X - a)Q + R,$$

avec $d^0(R) < d^0(X - a) = 1$; donc R est un polynôme constant. Prenant les valeurs des deux membres en a, on a P(a) = R, d'où l'égalité :

$$P = (X - a)Q + P(a). \qquad (9)$$

On dit que l'élément $a \in$ A est une *racine* du polynôme P si P(a) = 0. D'après (9), cela équivaut à avoir P divisible par le polynôme du premier degré X-u. On appelle *ordre de mulplicité* de la racine a le plus grand entier h tel que $(X - a)^h$ divise P ; ainsi, dire que a est racine d'ordre k du polynôme P équivaut à affirmer : $P = (X - a)^k Q$, l'élément a n'étant pas racine du polynôme Q.

Nous renvoyons aux articles CORPS et ÉQUATIONS ALGÉBRIQUES pour une étude détaillée des racines dans le cas où A est un corps commutatif. Terminons sur un résultat valable pour les anneaux d'intégrité : Si A est un anneau d'intégrité (unitaire) et si P $\in$ A[X] est de degré $\leqslant n$, P $\neq 0$, la somme des ordres de multiplicité de P dans A est $\leqslant$ n. Il en résulte que, si P et Q $\in$ A[X] sont tous deux de degré $\leqslant n$ et prennent des valeurs égales sur $n + 1$ éléments de A, alors P = Q.

Si K est un corps, on en déduit qu'il existe un polynôme de degré n et un seul prenant des valeurs données $b_i \in K$ sur n éléments distincts donnés $a_i \in K$, $i = 1, \dots, n$. On doit à Lagrange son expression sous la forme :

$$P(X) = \sum_{i=1}^{n} b_i R_i(X),$$

où $R_i(X)$ est le polynôme :

$$\frac{(X - a_1) \dots (X - a_{i-1})(X - a_{i+1}) \dots (X - a_n)}{(a_i - a_1) \dots (a_i - a_{i-1})(a_i - a_{i+1}) \dots (a_i - a_n)};$$

nous renvoyons à l'article représentation et approximation des FONCTIONS pour plus de précisions sur l'interpolation polynomiale. Remarquons que $R_i(X)$ est l'unique polynôme de degré n tel que :

$$R_i(a_j) = \delta_{ij},$$

où δ_{ij} est le symbole de Kronecker.

JEAN-LUC VERLEY

Bibliographie

L. CHAMBADAL & J.-L. OVAERT, *Cours de mathématiques*, t. II : *Algèbre II*, Gauthier-Villars, Paris. 1972 / R. GODEMENT, Cours *d'algèbre*, Hermann, Paris, 3' éd. 1994 / J. PICHON. *Les Polynômes*, Marketing, Paris, 1989.

POLYNÔMES ORTHOGONAUX → ORTHOGONAUX POLYNÔMES

POTENTIEL & FONCTIONS HARMONIQUES

La théorie du potentiel, directement issue de l'électrostatique, est une source d'inspiration extrêmement riche en analyse. Si, au début du XIXe siècle, on connaissait déjà l'équation de Laplace, la fonction de Green et l'intégrale de Poisson dans la boule, ce n'est vraiment qu'avec C. F. Gauss (1840) que sont posés et résolus, bien qu'imparfaitement, les grands problèmes de la théorie. Les idées de ce dernier sont si exemplaires qu'elles sont encore utilisées à l'heure actuelle, et il fallut attendre Frostman (1935) pour que le travail de Gauss fût amélioré en précision et en rigueur par l'introduction des outils nouveaux que s'était entre-temps forgés l'analyse.

On commencera par exposer les éléments fondamentaux qui permettront d'énoncer les problèmes et les principes les plus importants et les plus spécifiques de la théorie du potentiel : théorèmes de convergence. principe de domination, problème du balayage (cela afin d'arriver de la façon la plus directe à des résultats suffisamment précis). On se limitera ainsi au cas d'un ouvert borné, et on omettra de parler de « topologie fine » et de tout un chapitre de la théorie fine du potentiel. On ne pourra pas, non plus, parler des théories plus spéciales, comme par exemple la théorie très importante des potentiels besséliens d'Aronszajn et Smith. On envisagera quelques-unes des théories axiomatiques ou dérivées, issues de la théorie du potentiel, en mettant l'accent sur le lien avec la théorie des probabilités et celle des équations aux dérivées partielles : c'est là un centre de recherche important.

Le **sujet est** immense et la théorie du potentiel occupe une position centrale en analyse. En étudiant l'existence d'une solution du problème de Dirichlet, I. Fredholm considéra l'équation intégrale qui porte son nom ; c'est aussi à l'occasion de l'étude d'un critère de polyharmonicité que Laurent Schwartz a été amené à définir les distributions. Les problèmes de Dirichlet et de Neumann sont également des problèmes fondamentaux de la théorie des équations aux dérivées partielles. Enfin, la théorie de la capacité peut être considérée comme un chapitre important de la théorie de la mesure (ou vice versa), et le théorème de représentation intégrale de Choquet est tout aussi surprenant par sa simplicité que par sa profondeur et son efficacité.



1. Fonctions surharmoniques et potentiels

Pour tout ensemble A de $\mathbf{R}^n$, on note ∂A sa frontière topologique et $\bar{A}$ son adhérence. $B(x, r)$ désigne une boule ouverte de centre x et de rayon r.

La mesure de Lebesgue est notée dx et on entend par fonction réelle une fonction

à valeurs réelles prenant éventuellement les valeurs $+\infty$ et $-\infty$.

Fonctions surharmoniques et harmoniques

Une fonction réelle u définie dans un ouvert ω de $\mathbf{R}^n$, $n \geqslant 2$, est dite *hyperharmonique* si elle est semi-continue inférieurement et $> -\infty$, et si, pour tout $x \in \omega$ et pour toute boule $B = B(x, r)$, $\overline{B} \subset \omega$, on a :

$$u(x) \geqslant \frac{1}{\sigma(B)} \int_{\partial B} u \, d\sigma,$$

où $d\sigma$ désigne la mesure superficielle de la boule et $\sigma(B)$ l'aire de la boule B. On exprime cette dernière condition en disant que u majore sa moyenne sur toute boule.

De manière analogue, u est dite *hypoharmonique* si elle est semi-continue supérieurement et $< +\infty$ et si, avec les notations précédentes,

$$u(x) \leqslant \frac{1}{\sigma(B)} \int_{\partial B} u \, d\sigma,$$

pour $\overline{B} \subset \omega$.

Le module ou le logarithme du module d'une fonction holomorphe de la variable complexe z est hypoharmonique.

On dit qu'une fonction f définie dans un ouvert ω de $\mathbf{R}^n$ vérifie le « principe du minimum » si, pour tout ouvert δ, avec $\overline{\delta}$ compact $\subset \omega$, la condition :

$$\liminf_{y \to x} f(y) \geqslant \lambda,$$

pour tout $x \in \partial\delta$, entraîne $f \geqslant \lambda$ dans δ.

Propriétés des fonctions hyperharmoniques

1. Dans un ouvert ω, une fonction hyperharmonique ne peut atteindre un minimum en un point de ω sans être constante au voisinage. Cela entraîne que les fonctions hyperharmoniques dans un ouvert vérifient le principe du minimum.
2. L'ensemble des fonctions hyperharmoniques forme un cône convexe qui est stable par enveloppe inférieure finie.
3. L'enveloppe supérieure d'un ensemble filtrant croissant de fonctions hyperharmoniques est hyperharmonique.
4. Dans un domaine $\omega \subset \mathbf{R}^n$, une fonction hyperharmonique finie en un point est finie presque partout. Elle est alors dite *surharmonique.*
5. Une fonction localement surharmonique est surharmonique.
6. Si s est surharmonique dans un ouvert $\omega \subset \mathbf{R}^n$, pour tout p et tout compact $K \subset \omega$, il existe une suite croissante $\{s_n\}$ de fonctions surharmoniques p fois continûment différentiables dans un ouvert δ contenant K, telle que :

$$s(x) = \sup_n s_n(x),$$

pour tout $x \in \delta$.

Les propriétés 2, 3 et 4 sont évidentes. Les propriétés 1, 5 et 6, plus difficiles, sont fondamentales.

Une fonction u telle que $-u$ soit surharmonique est dite sous-harmonique. Si u est surharmonique et sous-harmonique, elle est dite *harmonique.* Elle est donc finie, continue et égale à sa moyenne en tout point.

Propriétés des fonctions harmoniques

1. Une fonction harmonique u dans un ouvert $\omega \subset \mathbf{R}^n$ ne peut avoir un maximum ou un minimum en un point de ω sans être constante au voisinage.
2. Les fonctions harmoniques sont indéfiniment dérivables.

Il suffit pour le voir de faire le produit de convolution de u harmonique dans un

ouvert ω avec la fonction indéfiniment dérivable φ définie par :

$$\varphi(x) = \begin{cases} k \exp \frac{-1}{r^2 - r_0^2}, & 0 \leqslant r < r_0, \\ 0, & r \geqslant r_0, \end{cases}$$

$r = \|x - t_0\|$, pour $\overline{B(t_0, r_0)} \subset \omega$, la constante k Etant choisie de telle sorte que :

$$\int_{B(t_0, r_0)} \varphi(x)\, dx = 1.$$

En écrivant u sous le signe $\int$ en fonction de sa valeur moyenne et en intervertissant l'ordre des intégrations, on vérifie que la fonction $\varphi * u$, définie par :

$$\varphi * u(t) = \int_B \varphi(t - x) u(x)\, dx,$$

est indéfiniment dérivable en t_0, et vaut $u(t)$ au voisinage de t_0.

3. Une fonction u est harmonique dans ω si et seulement si, pour toute boule $B \subset \overline{B} \subset \omega$, on a, en désignant par du/dn la dérivée ***normale*** de u sur dB.

$$(1) \qquad \int_{\partial B} \frac{du}{dn}\, d\sigma = 0,$$

ce qu'on exprime en disant que le flux sortant est nul.

En effet, si u est harmonique, la moyenne de u sur $B(t_0, \gamma)$ ne dépend pas de r et, en dérivant sous le signe $\int$, on obtient (1). Inversement, l'intégration de (1) en γ montre que u vaut sa moyenne.

4. Une fonction u est harmonique si et seulement si elle vérifie $\Delta u = 0$ où Δ désigne le laplacien.

Cela résulte immédiatement de la formule de Green :

$$(2) \qquad \iint_B \Delta g\, dx = \int_{\partial B} \frac{dg}{dn}\, d\sigma.$$

Ainsi, la partie réelle d'une fonction holomorphe est harmonique (cf. FONCTIONS ANALYTIQUES Fonctions analytiques d'une variable complexe, chap. 3).

5. Une fonction harmonique est analytique.

6. Dans le plan, l'inversion et les transformations conformes conservent l'harmonicité. Dans $\mathbf{R}^n$, $n \geqslant 3$, l'inversion ne conserve pas l'harmonicité : si une inversion I_{x_0} de centre x_0 transforme x en x', u harmonique se transforme en V tel que $V(x') = u(x)$, alors :

$$h(x') = \frac{V(x')}{|x_0 - x'|^{n-2}}$$

est harmonique. La fonction h s'appelle la transformée de Kelvin de u.

La transformation de Kelvin (inversion dans le plan) conserve également la surharmonicité. Ajoutons aussi qu'une fonction s surharmonique vérifie $\Delta s \leqslant 0$ au sens des distributions. Cela se voit facilement pour s de classe C^2 à l'aide d'un développement limité à l'ordre 2. On passe au cas général en utilisant une suite croissante d'après la propriété 6 des fonctions surharmoniques.

Potentiel newtonien et logarithmique

En écrivant le laplacien en coordonnées polaires et en cherchant des solutions de $\Delta u = 0$ qui ne sont fonction que de la distance r à l'origine, on trouve les fonctions :

$$a \ln r + b,$$

pour $n = 2$, et les fonctions :

$$\frac{a}{r^{n-2}} + b,$$

pour $n \geqslant 3$, a et b étant des constantes arbitraires,

On introduit alors la fonction harmonique fondamentale h(r). Dans le plan :

$$h(r) = \ln \frac{1}{r}$$

s'appelle *noyau logarithmique.* Dans l'espace $\mathbf{R}^n$, $n \geqslant 3$,

$$h(r) = \frac{1}{r^{n-2}}$$

est le *noyau newtonien.* La fonction h est surharmonique et harmonique en dehors de l'origine.

Si μ est une mesure à support compact (par exemple), la fonction :

$$U^\mu(x) = \int h(x-y)\, d\mu(y)$$

s'appelle *potentiel logarithmique ou newtonien* selon que l'on se place dans le plan ou l'espace. On parlera aussi de potentiel classique si l'on ne veut pas préciser : il faut noter que ce potentiel n'est pas partout défini. En pensant à une charge électrique répartie (par exemple continûment) avec une densité ρ sur une surface Σ, on a :

$$U^\rho(x) = \int h(x-y)\,\rho(y)\, d\sigma(y),$$

où do représente la mesure d'aire de la surface. C'est pourquoi on utilise un langage imagé en disant que la mesure μ représente la charge ou les masses du potentiel. Par dérivation sous le signe somme, on vérifie facilement que U^μ est harmonique en dehors du support de μ. Il faut aussi savoir que, si μ est positive, U^μ est surharmonique.

Il y a une différence essentielle entre le cas du plan et celui de l'espace, qui provient de la différence de comportement des noyaux à l'infini : toute fonction surharmonique positive dans $\mathbf{R}^2$ tout entier est constante.

Intégrale de Poisson et problème de Dirichlet

La formule de Poisson dans la boule B = B(O, R) :

$$(3) \quad I^B(x) = \frac{1}{\alpha_n R}\int_{\partial B} \frac{R^2 - |x|^2}{|x-y|^n}\, u(y)\, d\sigma(y),$$

où α_n désigne l'aire de la boule de rayon 1, peut s'obtenir de diverses manières. La plus simple consiste à utiliser la transformation de Kelvin de la façon suivante : Si x' désigne l'inverse de x dans l'inversion de centre 0 et de puissance R^2, on effectue la transformation de Kelvin de centre x' qui transforme x en 0 et qui conserve $\partial B(O, R)$. La fonction $u(y)$ devient alors $h(y)$, et on écrit que la valeur en 0 est la moyenne de h sur dB. En transformant cette intégrale sur h en intégrale relative à u, on obtient la relation cherchée.

La formule de Poisson permet de résoudre le problème de Dirichlet dans le cas de la boule : Sifest une fonction donnée finie continue sur dB, alors I_f^B est un prolongement continu de f dans $\overline{B}$, harmonique dans B.

Inégalités de Harnack et familles de fonctions harmoniques

Une majoration suivie d'une minoration de (3) donne, pour $u > 0$ harmonique dans B(O, R), les deux formules suivantes :

$$(4) \quad R^{n-2}\frac{R-|x-y|}{R+|x-y|^{n-1}} \leqslant \frac{u(x)}{u(y)} \leqslant \frac{R+|x-y|}{R-|x-y|^{n-1}} R^{n-2},$$

$$(5) \quad |\overrightarrow{\text{grad}}\, u|_0 \leqslant \frac{n}{R} u(O).$$

On en déduit immédiatement :

Théorème 1 : Toute famille $\mathfrak{F}$ localement bornée de fonctions harmoniques

dans un ouvert ω est équicontinue en tout point de ω.

Corollaire 2 *(Ascoli)* : La convergence simple d'une suite $\{u_n\}$ localement bornée entraîne la convergence uniforme locale vers une fonction harmonique. Si ω est connexe, il suffit, grâce à l'analyticité, que $\{u_n\}$ converge simplement au voisinage d'un point.

De toute suite $\{u_n\}$ on peut extraire une suite $\{u_{np}\}$ convergeant uniformément localement vers une fonction harmonique.

L'inégalité (5) montre encore que la convergence uniforme locale de u entraîne celle des dérivées successives.

En partant de (4) ou du théorème précédent, on peut montrer facilement les deux propriétés suivantes, d'ailleurs équivalentes dans un domaine ω :

a) Pour tout K compact inclus dans ω, il existe une constante $k > 0$ dépendant seulement de K et de ω telle que, pour toute fonction harmonique $u > 0$ dans ω, on ait les inégalités de Harnack :

$$\frac{1}{k} \leqslant \frac{u(x)}{u(y)} \leqslant k$$

quels que soient x, y $\in$ K.

b) Pour toute famille (u,) filtrante croissante de fonctions harmoniques dans ω, si sup u_i est finie en un point, sup u_i est finie et harmonique dans ω.

L'équicontinuité jointe à l'inégalité de Harnack permet d'énoncer que toute famille (u,) de fonctions harmoniques, localement bornée inférieurement dans un domaine, forme une famille normale de Monte1 ; autrement dit : de toute suite (u,,) C (u_i) on peut extraire une suite convergeant vers $+\infty$ ou vers une fonction harmonique (uniformément localement).

Théorème de représentation de Riesz

Soit y un point de $B(x_0, R)$. On appelle *fonction de Green* ou bien *noyau de Green* de la boule $B(x_0, R)$ relative au pôle y la fonction :

$$G_y^B = h_y - I(h_y).$$

où $I(h_y)$ est l'intégrale de Poisson dans B de la restriction de h_y à dB.

Si y, est l'inverse de y dans l'inversion de centre x_0 qui conserve dB, on a explicitement :

$$G_y(x) = h(|x-y|) - h\left(\frac{|y-x_0| \cdot |x-y_1|}{R}\right).$$

La fonction G_y, surharmonique positive dans B, est harmonique en dehors de y, s'annule sur dB et vérifie la propriété de symétrie $G_y(x) = G_x(y)$.

On définirait plus généralement la *fonction de Green* d'un ouvert borné ω en posant :

$$G_y^\omega = h_y - H_{h_y}^\omega,$$

où $H^\omega_{h_y}$ désigne la solution généralisée du problème de Dirichlet avec pour donnée frontière la trace de h_y sur $\partial\omega$. La seule différence provient de ce que G_y ne s'annule pas partout à la frontière.

Si μ est une mesure positive, la fonction :

$$G_\mu^\omega(x) = \int G_y^\omega(x)\, d\mu(y)$$

est appelée potentiel de Green, potentiel pur ou simplement potentiel. C'est une fonction hyperharmonique positive dans ω.

Si le potentiel U^μ est deux fois continûment différentiable et $\mu = \rho\, dx$, avec ρ continue, on obtient, à l'aide de la formule de Green, l'équation de Poisson :

$$\Delta U^\mu = -\varphi_n \rho,$$

où φ_n est un coefficient numérique dépendant de n.

Plus généralement, si T est une distribution à support compact, U^T se définit encore comme la distribution $h * T$, et on a encore au sens des distributions :

$$AU= = - \varphi_n T$$

L'équation de Poisson permet de connaître la charge quand on connaît le potentiel et permet de démontrer le théorème suivant.

Théorème 3 (théorème de représentation de Riesz) : Si V, surharmonique dans un ouvert borné ω, admet une minorante harmonique dans ω, on a :

$$V(x) = \int G_y^{\omega}(x)\, d\mu(y) + V^*(x),$$

où V* est la plus grande minorante harmonique de V et μ la mesure positive $\Delta V/\varphi_n$.

L'existence d'une plus grande minorante harmonique résulte de la proposition suivante.

Proposition 4 : Si 3 est la famille des fonctions sous-harmoniques minorant la fonction surharmonique V, alors h = sup 3 est harmonique.

La fonction h est alors la plus grande minorante harmonique de V. Cela se démontre en utilisant le fait que l'on peut remplacer s sous-harmonique dans ω par la fonction s' égale à I_s^B dans la boule B C $\overline{B}$ C ω et égale à s ailleurs. De ce fait, $s' \leqslant s$, et s' est encore sous-harmonique. On remarque que $\mathcal{F}$ est filtrante croissante. Si $\mathcal{F}'$ désigne la famille des fonctions s' quand s parcourt 3 pour B donné, on voit que sup $\mathcal{F}$ = sup 3' est harmonique dans B (théorème de convergence pour les fonctions harmoniques) ; B étant arbitraire, h est harmonique (critère local).

Notons qu'un potentiel est donc caractérisé comme une fonction surharmonique positive dont la plus grande minorante harmonique est nulle.

2. Théorèmes et principes fondamentaux

Balayage

On appelle S-fonction une fonction u localement bornée inférieurement qui vérifie, pour toute boule $B(x, R)$, la relation :

$$u(y) \geqslant \frac{1}{\sigma(B)} \int_{\partial B}^{*} u\, d\sigma, \quad y \in B(x),$$

où $d\sigma$ est la mesure-aire de dB et $\int^*$ l'intégrale supérieure.

L'utilité de ces fonctions provient de ce que l'enveloppe inférieure d'une famille de S-fonctions localement bornée inférieurement est une S-fonction et que la régularisée semi-continue inférieurement d'une S-fonction est hyperharmonique.

Soit ω un ouvert borné de $\mathbf{R}^n$, E C ω, et φ une fonction $\geqslant 0$ sur E. On note $(R_\varphi^E)_\omega$ ou R_φ^E l'enveloppe inférieure des fonctions v hyperharmoniques $\geqslant 0$ dans ω qui majorent φ sur E. La fonction R_φ^E s'appelle la *réduite* de φ sur E et est une S-fonction ; c'est une fonction croissante de E, positivement homogène et sous-additive en φ.

La proposition 4 montre que R_φ^E est harmonique, ou égale à + ∞, en dehors de E.

Si φ est la trace d'une fonction surharmonique $v \geqslant 0$, la régularisée $\hat{R}_v^E$, alors surharmonique, est appelée la *balayée* de v sur E. Si E est suffisamment régulier (une boule, par exemple), la balayée vaut v sur E ; et, si v est un potentiel G_μ, la balayée,

majorée par G_μ, est encore un potentiel qui vaut G_μ sur E. On peut donc écrire :

$$R^E_{G_\mu} = G_{\mu_E}.$$

On dit que μ_E est la balayée de μ sur E et que μ_E engendre le même potentiel que G_μ sur E. On dit de façon imagée que l'on a balayé les masses sur E. C'est en fait ce qui se passe : si E est toujours un compact suffisamment régulier, μ_E est alors constituée des masses de μ portées par E, auxquelles viennent s'ajouter les masses de μ qui n'étaient pas portées par E. Une étude approfondie donne un résultat plus précis pour E quelconque.

Principe de domination

Plusieurs formes plus ou moins fortes du principe de domination peuvent être données, dont celle-ci : Si, dans un ouvert ω, une fonction surharmonique majore un potentiel G,, localement borné sur le support de μ, elle le majore partout. Si G,, était continu, il ne s'agirait de rien d'autre que du principe du minimum.

Capacité

Soit K un compact d'un ouvert borné ω de R". On appelle *potentiel capacitaire* de K la balayée $\hat{R}^K_1$ de 1 sur K. La mesure qui l'engendre s'appelle la *mesure capacitaire* et p(K) est la *capacité* de K, notée C(K). Grâce au principe de domination, on voit que $\hat{R}^K_1$ est le plus grand potentiel majoré par 1 dans ω ayant une masse associée $\geqslant 0$ portée par K. Ce résultat donne la formule plus connue :

$$\mathcal{C}(K) = \sup \mu(K),$$

où la borne supérieure est prise pour l'ensemble de toutes les mesures positives μ portées par K telles que $G_\mu \leqslant 1$ sur K.

La capacité $\mathcal{C}$ est une fonction d'ensemble définie sur les compacts de ω et telle que :
1. La fonction $\mathcal{C}$ est une fonction croissante :

$$K_1 \subset K_2 \Rightarrow \mathcal{C}(K_1) \leqslant \mathcal{C}(K_2);$$

2. Si (K,,) est une suite décroissante de compacts, d'intersection K $\neq \varnothing$, on a :

$$C(K) = \lim_n \mathcal{C}(K_n);$$

on dit que $\mathcal{C}$ descend sur les compacts ;
3. On a la propriété de sous-additivité forte. Pour tous compacts K, et K_2, on a :

$$\mathcal{C}(K_1 \cup K_2) + \mathcal{C}(K_1 \cap K_2) \leqslant \mathcal{C}(K_1) + \mathcal{C}(K_2).$$

On définit aussi la capacité intérieure $\mathcal{C}_*$ (E) et la capacité extérieure C*(E) d'un ensemble quelconque E par :

$$\mathcal{C}_*(E) = \sup \mathcal{C}(K),$$

où la borne supérieure est prise pour K parcourant l'ensemble des compacts contenus dans E, et :

$$\mathcal{C}^*(E) = \inf \mathcal{C}_*(\omega),$$

où la borne inférieure est prise pour ω parcourant l'ensemble des ouverts contenant E.

On montre que, pour E relativement compact dans ω, C*(E) est encore la masse totale de la mesure associée à $\hat{R}^E_1$.

La fonction $\mathcal{C}^*$ est croissante, descend sur les compacts et monte sur les ensembles quelconques ; cela signifie que, si (E_n) est une suite croissante d'ensembles quelconques. on a :

$$\mathcal{C}^*(\bigcup_n E_n) = \sup_n \mathcal{C}^*(E_n).$$

Ensembles exceptionnels

Il existe des ensembles de capacité strictement positive, qui sont très petits. Par

exemple, dans le plan, l'ensemble fermé de Cantor sur le segment [0, 1] est de capacité strictement positive alors qu'il est non seulement de mesure nulle (dans $\mathbf{R}^2$), mais de mesure linéaire nulle. La capacité mesure donc de façon très fine la petitesse des ensembles. On verra qu'elle est parfaitement adaptée à la mesure des ensembles exceptionnels de la théorie du potentiel.

On dit qu'une propriété est vraie quasi partout (on note en abrégé q.p.) si elle est vraie en dehors d'un ensemble de capacité extérieure nulle. Un ensemble P est dit *polaire* dans un ouvert borné s'il existe un potentiel G_μ fini en un point et infini sur P (contrairement aux apparences, cela ne dépend pas de ω borné).

H. Cartan a montré qu'il y avait identité entre ensemble polaire et ensemble de capacité extérieure nulle dans un ouvert borné. Il a aussi démontré le théorème suivant, qui avait été obtenu précédemment par M. Brelot pour la capacité intérieure. Ce théorème est la clef de toutes les études fines de la théorie.

Théorème 5 (théorème de Cartan-Brelot) : L'enveloppe inférieure d'une famille (v_i) localement bornée inférieurement de fonctions surharmoniques dans un ouvert ω de $\mathbf{R}^n$ diffère de sa régularisée semi-continue inférieurement d'un ensemble polaire. Cela permet de préciser pour un ensemble E quelconque que $\hat{R}_v^E$, pour v surharmonique $\geqslant 0$, est égale q.p. à v sur E.

Problème de Dirichlet généralisé et effilement

Soit ω un ouvert borné et f une donnée frontière (finie ou non). On considère la famille Φ des fonctions hyperharmoniques v telles que :

$$\liminf_{\substack{x \to y \\ x \in \omega}} v(x) \geqslant f(y),$$

pour tout $y \in$ do, cette limite inférieure étant > $-\infty$ partout sur $\partial\omega$. On définit :

$$\overline{H}_f^\omega = \inf \Phi, \ \underline{H}_f = -\overline{H}_{(-f)} ;$$

on a alors :

$$\underline{H}_f \leqslant \overline{H}_f,$$

et $\overline{H}_f(x)$ est une fonction positivement homogène et sous-additive de f. Si on a $\underline{H}_f = \overline{H}_f = H_f$, on dit que f est *résolutive* et H, est alors la solution généralisée du problème de Dirichlet. Cette construction est due à Perron. Pour sa part, N. Wiener montra queffinie continue était résolutive et qu'ainsi $f \mapsto H_f(x)$ définissait une mesure de Radon positive ρ_x^ω appelée *mesure harmonique*. Brelot montra alors que pour qu'une fonction f, définie sur do, soit résolutive, il faut et il suffit qu'elle soit ρ_x^ω intégrable ; cela ne dépend pas de XE ω.

Un point frontière $x_0 \in$ do est dit *régulier* si, pour toutef'finie continue sur do, on a :

$$\lim_{\substack{x \to x_0 \\ x \in \omega}} H_f(x) = f(x_0) ;$$

si tout point frontière est régulier, on dit que ω est *régulier*. C'est le cas de la boule.

On dit qu'un ensemble E est *effilé* en un point $x_0 \notin$ E si x_0 n'est pas adhérent à E ou s'il existe une fonction surharmonique v au voisinage de E telle que :

$$v(x_0) < \liminf_{\substack{x \to x_0 \\ x \in E}} v(x) ;$$

si $x_0 \in$ E, on dit encore que E est *effilé* en x_0 si E $- \{x_0\}$ est effilé en x_0.

On montre alors qu'il existe un potentiel fini continu P dans ω (ouvert borné) tel que, pour tout $E \subset \omega$, l'ensemble des points d'effilement de E coïncide avec l'ensemble des points x pour lesquels on a :

$$\widehat{R}_P^E(x) < P(x).$$

On en déduit, d'après le théorème 5, que l'ensemble des points de E, où E est effilé, est polaire (ou de capacité extérieure nulle). Si ω est un ouvert borné, on montre que x_0 est régulier pour ω si et seulement si son complémentaire est non effilé en x.

Il s'ensuit que l'ensemble des points irréguliers forme un ensemble polaire.

Fonctions harmoniques positives et frontières de Martin

Toute fonction harmonique positive u dans la boule B(O,R) s'écrit :

$$u(x) = \frac{1}{\alpha_n R} \int \frac{R^2 - |x|^2}{|x-y|^n} \, d\mu(y),$$

où μ est une certaine mesure $\geqslant 0$ sur dB. Citons aussi le théorème de Fatou qui affirme que toute fonction harmonique positive dans la boule B admet une limite angulaire en presque tout point $x \in$ dB.

En 1941, R. S. Martin, afin de généraliser cette représentation intégrale au cas d'un ouvert borné, introduisit la fonction *de Green normalisée* :

$$K(x,y) = \frac{G^{\omega}(x,y)}{G^{\omega}(x,y_0)}, \quad y_0 \in \omega ;$$

il montra l'existence d'un espace compact $\hat{\omega}$, unique à un homéomorphisme près, tel que $\hat{\omega} = \overline{\omega}$ et prouva que la famille des fonctions $x \mapsto K(x, y)$ se prolonge continûment à $\Gamma = \hat{\omega} - \omega$ et sépare Γ. On appelle Γ_1 l'ensemble des points $X \in \Gamma$ tels que la fonction prolongée correspondante $K(X, y)$ soit minimale, c'est-à-dire telle que toute fonction harmonique > 0 majorée par $K(X, y)$ lui soit proportionnelle. À toute fonction harmonique $h \geqslant 0$ dans ω correspond alors sur Γ une mesure unique $\mu_h \geqslant 0$ portée par Γ_1 telle que :

$$h(y) = \int K(x,y) \, d\mu_h(x).$$

Brelot a résolu un problème de Dirichlet avec donnée frontière sur Γ. On peut alors faire une étude du comportement à la frontière des fonctions harmoniques positives, grâce à l'introduction de l'effilement minimal (Naïm), et généraliser, à l'aide de la topologie fine, le résultat de Fatou sur les limites angulaires dans le cas de la boule (J. L. Doob).

3. Liens avec l'analyse fonctionnelle

Énergie

La physique élémentaire nous apprend que l'unique charge électrique q du potentiel capacitaire V d'un conducteur donne un état d'équilibre et correspond à un minimum de l'énergie :

$$\frac{1}{2} \int V \, dq ;$$

on dit que V est un potentiel d'équilibre. C'est cette idée qui conduisit Gauss, en 1840, à considérer l'intégrale :

$$\int (U^{\mu} - 2f) \, d\mu,$$

où U^{μ} est le potentiel newtonien d'une mesure μ donnée par une densité sur une surface Σ rendant minimum l'intégrale. Or cela n'est vrai qu'avec des restrictions qui furent éclaircies par Frostman en 1935. Ce sont, au fond, les idées de Gauss qui sont à l'origine du travail de Cartan sur l'énergie dont il est question ci-dessous.

Dans $\mathbf{R}^n$, $n \geqslant 3$, avec le noyau newtonien, on appelle *énergie mutuelle* de deux mesures μ et $\nu \geqslant 0$ la quantité :

$$(\mu|\nu) = \int U^\mu d\nu = \int U^\nu d\mu ;$$

pour toute mesure μ, on appelle *énergie* de μ le nombre $\|\mu\|_e = \sqrt{(\mu|\mu)}$ et, à l'aide de l'inégalité fondamentale (non évidente),

$$(\mu\nu) \leqslant \|\mu\|_e \|\nu\|_e,$$

il est facile de voir que $\mu \mapsto \|\mu\|_e$ est une semi-norme et, par suite, que l'ensemble des mesures positives ou nulles d'énergie finie est un cône convexe $\mathcal{E}^+$.

On considère ensuite l'espace vectoriel $\mathcal{E} = \mathcal{E}^+ \quad \mathcal{E}^+$ et on prolonge de façon standard la semi-norme à $\mathcal{E}$. L'inégalité fondamentale est encore vérifiée et la semi-norme prolongée est encore une semi-norme sur $\mathcal{E}$.

Principaux théorèmes

Théorème 6 (principe de l'énergie de Frostman) : La semi-norme $\|\mu\|_e$ est une norme sur $\mathcal{E}$.

Corollaire 7 : L'espace $\mathcal{E}$ est un espace préhilbertien muni du produit scalaire $(\mu \ \nu)$.

Théorème 8 (principe de domination ou principe du maximum de Cartan) : Soit $\mu \in \mathcal{E}$ et v surharmonique positive majorant U^μ sur un support restreint de u (c'est-à-dire un ensemble dont le complémentaire est de u-mesure nulle) ; alors v majore U^μ partout dans $\mathbf{R}^n$.

Théorème fondamental (Cartan) : Le cône convexe $\mathcal{E}^+$ est complet dans $\mathcal{E}$.

On utilise pour la démonstration de ce théorème le théorème de représentation de Riesz (théorème 3). Un exemple de Cartan montre que $\mathcal{E}$ n'est pas complet.

Si K est un compact de $\mathbf{R}^n$, on peut montrer que l'ensemble $\mathcal{F}_K$ des mesures A $\in \mathcal{E}^+$ portées par K est un cône convexe complet de $\mathcal{E}$; on obtient (9) et (10).

Théorème 9 (du balayage) : Soit $\mu \in \mathcal{E}^+$. La projection μ_K de μ sur $\mathcal{F}_K$ est caractérisée comme la seule mesure $\geqslant 0$ sur K telle que, d'une part, on a partout l'inégalité :

$$U^{\mu_K} \leqslant U^\mu$$

et, d'autre part, pour toute mesure A $\in \mathcal{E}^+$, on a, h-presque partout,

$$U^{\mu_K} = U^\mu.$$

On voit que μ_K est la balayée de μ grâce à la remarque qui suit le théorème 5 et au théorème suivant.

Théorème 10 : Un borélien de $\mathbf{R}^n$ est polaire si et seulement s'il est de A-mesure nulle pour toute mesure A $\in \mathcal{E}^+$.

Norme et principe de Dirichlet

Soit $\mathcal{T}_0$ l'espace des fonctions numériques possédant un gradient fini continu de carré intégrable. Pour toute $u \in \mathcal{T}_0$, on pose :

$$\|u\| = \left(\int \sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_i} \right)^2 dx \right)^{1/2}.$$

L'application $u \mapsto \|u\|$ est une semi-norme dans $\mathcal{T}_0$ associée au produit scalaire :

$$(u_1, u_2) = \int \overrightarrow{\text{grad}}\, u_1 \cdot \overrightarrow{\text{grad}}\, u_2 \, dx ;$$

la condition $\|u\| = 0$ équivaut à u constante.

Pour obtenir une norme on passe au quotient $\mathcal{T}$ par la relation d'équivalence naturelle. La norme correspondante s'appelle la *norme de Dirichlet*. Par commodité de langage, on confond une fonction de $\mathcal{T}_0$ avec sa classe d'équivalence dans Y.

Théorème 11 : Le sous-espace $\mathcal{H}$ des fonctions harmoniques est complet dans $\mathcal{T}$.

Cela permet d'énoncer le principe *de Dirichlet* : pour toute fonction $f \in \mathcal{F}$, il existe une fonction harmonique $u \in \mathcal{H}$ unique rendant minimum le nombre $\|u-f\|$.

En effet, $\mathcal{H}$ étant un sous-espace complet, on obtient u par projection.

On peut ainsi résoudre le *problème de Dirichlet* dans les conditions suivantes : Soit ω un domaine borné de $\mathbf{R}^n$, $n \geqslant 2$, et f une fonction de $\mathcal{F}_0$ bornée dans ω et admettant un prolongement fini et continu à $\overline{\omega}$ (noté encore f). La projection de f sur $\mathcal{H}$ est H_f^ω. Si, de plus, ω est un domaine régulier, H_f^ω est l'unique fonction de $\mathcal{F}_0$ de norme minimale admettant en tout point la même limite que f.

En songeant à compléter l'espace $\mathcal{F}$, J. Deny a été amené à introduire des fonctions appelées « BL précisées » ou « BLD », obtenues en précisant des fonctions introduites par Beppo Levi et Nikodym. Les classes d'équivalence de ces fonctions, par rapport à l'égalité presque partout à une constante près, forment un espace de Hilbert. On peut résoudre un problème de Dirichlet correspondant pour ces fonctions.

On peut faire le lien des fonctions BLD avec les potentiels U^μ d'énergie finie : Si $\mu \in \mathcal{E}$, alors U^μ est une fonction BLD et on a :

$$\|U^\mu\| = \varphi_n \|\mu\|_e,$$

où φ_n est un coefficient numérique dépendant de n.

4. Théories axiomatiques et dérivées

Méthodes hilbertiennes

L'espace $\mathcal{E}$ des mesures d'énergie finie n'étant pas complet, Deny, en 1950, introduit les éléments du complété en développant une théorie du potentiel dans $\mathbf{R}^n$, où le noyau est une distribution et le potentiel un produit de convolution de distributions (cf. DISTRIBUTIONS). Avec quelques restrictions, la théorie de Cartan peut être adaptée. Dans l'axiomatisation par Beurling et Deny des espaces de Dirichlet, on utilise le fait que la norme de Dirichlet est diminuée par les contractions normales : Si v varie moins vite que u, l'intégrale de Dirichlet relative à v est plus petite que l'intégrale de Dirichlet relative à u. Cette remarque, due à A. Beurling, permet de donner des démonstrations très courtes et très élégantes des résultats fondamentaux de la théorie du potentiel. Elle permet aussi de démontrer des théorèmes profonds de synthèse spectrale en analyse harmonique.

Théories axiomatiques sans noyaux

La théorie axiomatique de Brelot et de ses extensions ultérieures (Bauer, Constantinescu et Cornea) est inspirée d'une axiomatique probabiliste de Doob et fut précédée d'une tentative due à Tautz. Le principe en est le suivant.

Dans un espace Ω localement compact, on considère un faisceau d'espace vectoriel de fonctions numériques continues, appelées harmoniques (axiome 1). On suppose qu'il existe une base de domaines réguliers, c'est-à-dire tels qu'il existe une solution du problème de Dirichlet (axiome 2), et enfin (axiome 3) que tout ensemble filtrant croissant de fonctions harmoniques dans un domaine ω tend vers $+\infty$ ou une fonction harmonique. La théorie se développe considérablement si l'on ajoute le principe de domination (axiome D) comme nouvel axiome.

Les solutions dans un ouvert d'une équation du deuxième ordre de type elliptique à coefficients suffisamment réguliers

vérifient les axiomes. Il en est de même, ce qui est plus difficile, pour des équations à coefficients discontinus (Mme Hervé). Cela apporte de considérables simplifications à l'étude directe de ces équations faites par Stampacchia.

En revanche, les solutions d'équations de type parabolique ne vérifient pas les axiomes 3 et D. C'est pourquoi Bauer modifia l'axiomatique précédente par l'introduction d'un nouvel axiome et l'affaiblissement de l'axiome 3, afin de contenir, dans les applications, les solutions d'équations de ce type.

Enfin, J. M. Bony est arrivé à caractériser de façon presque complète, en termes d'opérateurs différentiels, les théories axiomatiques du type Brelot, Bauer... dans R".

Soit, par exemple, $\mathcal{H}$ une théorie axiomatique de Brelot telle qu'il y ait « suffisamment » de fonctions de classe $\mathcal{C}^\infty$. Il existe alors un ouvert dense dans lequel est défini un opérateur différentiel L à coefficients de classe $\mathcal{C}^\infty$ tel que toute fonction harmonique de classe $\mathcal{C}^2$ vérifie Lu = 0 et même encore, au sens des distributions, si u n'est pas de classe $\mathcal{C}^2$.

Théorie de Hunt et probabilités

Soit Ω un espace localement compact. On appelle noyau N une famille $\{\mu_x\}$ de mesures dépendant mesurablement (en un sens à préciser) de x. On note u,(E) = $N(x, E)$. À toute f borélienne $\geqslant 0$, on associe :

$$Nf_x = \int f\,d\mu_x = \int N(x, dy) f(y),$$

également borélienne. On peut considérer N comme une application linéaire positive de l'ensemble des fonctions boréliennes positives dans lui-même. On peut donc composer deux noyaux. Si v est une mesure positive, on définit la mesure νN par :

$$\nu N(E) = \int N(x, E)\, d\nu(x).$$

L'exemple classique est le noyau :

$$N(y, E) = \int_E h(|x - y|)\, dx$$

dans R". Ici Nfest le potentiel newtonien de densitéfet νN est absolument continue par rapport à la mesure de Lebesgue, ayant comme densité le potentiel U".

On dit que N satisfait au *principe complet du maximum* si, pour toute constante $a \geqslant 0$ et tout couple $(f$, g) de fonctions positives universellement mesurables, la relation :

$$a + Nf_x \geqslant Ng_x,$$

pour tout x tel que g(x) > 0, entraîne :

$$a + Nf_x \geqslant Ng_x,$$

pour tout $x \in \Omega$.

Avec certaines restrictions satisfaites dans les applications, G. A. Hunt montre qu'on peut associer à un noyau N, satisfaisant **au** principe complet du maximum, un semi-groupe P, (défini pour $t \geqslant 0$, $P_{s+t} = P_s \circ P_t$) de noyaux, vérifiant des conditions de continuité à l'origine, tel que :

$$Nf = \int_0^{+\infty} P_t f\, dt\,;$$

cela lui permet de développer une théorie du potentiel purement probabiliste. On appelle *excessives* les fonctions $f \geqslant 0$ vérifiant :

$$P_t f \leqslant f, \quad \lim_{t \to 0} P_t f = f\,;$$

en cas d'égalité, f est dite *invariante*. Quand $f \geqslant 0$ n'a pas de minorante invariante $\geqslant 0$ autre que zéro, f est appelée *potentiel*.

Dans les bons cas, P, peut être interprété comme le semi-groupe de transition d'un processus de Markov. De tels processus sont appelés processus de Hunt. Dans le cas particulier du mouvement brownien, le générateur infinitésimal (dérivée à l'origine du semi-groupe) est l'opérateur Δ, ce qui permet d'identifier les fonctions surharmoniques et les fonctions excessives. Il en est de même dans le cas général d'une théorie axiomatique du type Brelot. M[me] Hervé a construit un noyau vérifiant des conditions qui permirent à P. A. Meyer de montrer l'existence d'un semi-groupe dont les fonctions excessives sont précisément les fonctions surharmoniques de la théorie.

Cette identification va beaucoup plus loin et permet d'interpréter en termes probabilistes les faits les plus importants de la théorie du potentiel : balayage, effilement, espace de Martin, etc. La théorie du potentiel est donc une source d'inspiration considérable pour les probabilistes qui s'occupent des processus markoviens.

Théorème de représentation intégrale

Les fonctions harmoniques positives dans un ouvert borné $\omega \subset \mathbf{R}^n$ forment un cône convexe C : celles qui valent 1 en un point forment une base convexe compacte B (pour la convergence compacte) du cône et les fonctions minimales de cette base sont les éléments extrémaux de B. On peut aussi interpréter la représentation intégrale de Martin d'une fonction $u \in B$ en disant que u est le barycentre d'une mesure μ portée par l'ensemble des points extrémaux (modulo une identification des éléments minimaux, c'est-à-dire extrémaux, de B avec les points de l'ensemble Γ_1 défini à la fin du chapitre 2).

Cette remarque a permis à G. Choquet de démontrer le théorème extrêmement profond qui suit.

Théorème. Soit C un cône convexe et B une base compacte de C. Si B est métrisable, tout $y \in$ B est barycentre d'une mesure unitaire μ portée par l'ensemble des points extrémaux. De plus, si C est réticulé pour son ordre, μ est unique.

Si B n'est pas métrisable, le problème est beaucoup plus compliqué. En particulier, l'ensemble des points extrémaux n'est pas nécessairement mesurable.

On peut partir de ce théorème pour retrouver la représentation intégrale de Martin : c'est une méthode beaucoup plus simple. Ce théorème permet également de donner une représentation intégrale de Riesz dans les espaces harmoniques de Brelot (M[me] Hervé, G. Mokobodski) et même, sous une forme moins satisfaisante, dans les axiomatiques affaiblies (Mokobodski).

Théorie de la capacité

Une *capacité généralisée*, au sens de Choquet, sur un espace topologique séparé X est une fonction réelle $\mathcal{C}$ d'ensemble, définie sur toutes les parties de X ; elle est croissante, descend sur les compacts et monte sur les ensembles quelconques. La capacité extérieure classique dans $\mathbf{R}^n$ et les mesures extérieures sur X localement compact sont des exemples de capacité généralisée. Un ensemble $A \subset X$ est dit *capacitable* si :

$$\mathcal{C}(A) = \sup_K \mathcal{C}(K),$$

la borne supérieure étant prise pour K parcourant l'ensemble des compacts contenus dans A.

Pour terminer, indiquons un théorème, dû à Choquet, qui est très utile en théorie de la mesure et en théorie des probabilités.

Il nous faut pour cela donner quelques définitions : On dit qu'un ensemble dans un espace topologique est un K_σ si c'est une réunion dénombrable d'ensembles compacts ; un $K_{\sigma\delta}$ est un ensemble qui est intersection dénombrable de K_σ ; enfin, on dit qu'un sous-ensemble A d'un espace topologique séparé est *analytique* si A est l'image continue d'un $K_{\sigma\delta}$ contenu dans un espace compact. Le théorème de Choquet s'énonce alors : Tout ensemble analytique contenu dans un K_σ est capacitable.

ARNAUD DE LA PRADELLE

Bibliographie

S. J. AXLER, P. BOURDON & W. RAMEY, *Harmonic Function Theory*, Springer, New York, 1992 / M. BRELOT, *Élément de la théorie classique du potentiel*, C.D.U., Paris, 1959 ; *Séminaire de théorie du potentiel* 1972. univ. Pierre-et-Marie-Curie. Paris. 1973 ; *On Topologies and Boundaries in Potential Theory*, Springer-Verlag, Berlin, 1971 / J. KRAL et al., *Potentiel Theory*, Plenum Press, 1988 / N. LANDKOF, *Foundations of Modern Potential Theory*, Springer, Berlin, 1972 / P. A. MEYER & C. DELLACHERIE, *Probabilités et potentiel*, 5 vol., Hermann, Paris, 1976-1992.

PRODUITS INFINIS → SÉRIES & PRODUITS INFINIS

PROJECTIFS ESPACE & REPÈRE

Espace projectif. Étant donné un espace vectoriel E sur un corps commutatif K, on considère dans E' = E {0} la relation G entre deux éléments x et y définie par :

$$\exists\lambda \in K, \quad \lambda \neq 0, \quad y = \lambda x.$$

La relation G est une relation d'équivalence et l'ensemble quotient E'/G est appelé espace projectif déduit de E et est noté P(E). L'ensemble E est appelé espace vectoriel sous-jacent de P(E). Une classe d'équivalence, élément de P(E), est appelée point projectif; on désigne par π l'application canonique qui à un élément de E' associe sa classe dans P(E). Lorsque E = K^{n+1}, l'espace projectif déduit se note P,,(K). Si E est de dimension n + 1, la dimension de P(E) est, par définition, n. Il faut toutefois remarquer que P(E) n'est pas un espace vectoriel.

L'espace projectif réel ou complexe P,(R) ou P,,(C) est une variété compacte non orientable. L'expace affine réel ou complexe de dimension n se plonge de manière naturelle dans cet espace projectif ; ce plongement correspond géométriquement à l'adjonction de « points à l'infini », réels ou imaginaires, à cet espace affine.

Variété linéaire projective. Soit F un sous-espace vectoriel de E, l'image par π de F' = F {0} est, par définition. une variété linéaire projective de P(E). On peut aisément montrer **que** l'intersection d'une famille quelconque de variétés linéaires projectives est une variété linéaire projective et que l'espace vectoriel sous-jacent de cette intersection est l'intersection des espaces vectoriels sous-jacents des variétés de la famille. Une variété projective déduite d'un hyperplan de E s'appelle un hyperplan projectif, et sa dimension (lorsque dim (E) = $n + 1$) est égale à n 1 ; un espace projectif de dimension 1 (resp. 2) est appelé droite projective (resp. plan projectif). Soit X un sous-ensemble de P(E) ; on appelle variété linéaire engendrée par X l'intersection de

toutes les variétés linéaires contenant X. Soit $k+1$ points de P(E) ; on dit qu'ils forment une partie projectivement libre si la dimension de la variété engendrée par eux est égale à k ; ils sont projectivement liés si la dimension de la variété est inférieure à k. On peut montrer que $k+1$ points $\pi(x_i)$ de P(E) sont libres si et seulement si les $k+1$ points x_i sont libres dans E. Ainsi, bien que P(E) ne soit pas un espace vectoriel, la notion d'indépendance se conserve. Par suite, on a des énoncés de théorèmes sur les dimensions équivalents aux énoncés sur les dimensions des sous-espaces vectoriels, en particulier le théorème de la « base incomplète ».

Coordonnées homogènes ; repère projectif. Soit $B = (e_i)$, $1 \leqslant i \leqslant n+1$, une base de l'espace vectoriel E de dimension $n+1$. Tout élément x de E s'écrit :

$$x = \sum_{i=1}^{n+1} x_i e_i.$$

avec $x_i \in$ K. Le $(n+1)$-uple $(x_1, x_2, ..., x_{n+1})$ s'appelle système de coordonnées homogènes du point $\pi(x)$ de P(E). Soit e_0 l'élément de E de coordonnées :

$$e_0 = \sum_{i=1}^{n+1} e_i ;$$

on appelle repère projectif l'ensemble des $n+2$ points $\pi(e_0)$, $\pi(e_1, \ldots \pi(e_{n+1})$.

On peut donner une représentation, à l'aide de coordonnées, d'une variété linéaire projective P(F) : il suffit de donner la représentation du sous-espace vectoriel sous-jacent F privé de 0. Soit $\pi(g_j)$, $1 \leqslant j \leqslant k+1$, une famille libre engendrant P(F). Désignons par (a_{ij}), $1 \leqslant i \leqslant n+1$, le système de coordonnées homogènes du point $\pi(g_i)$. Alors un système de coordonnées homogènes d'un point de P(F) est donné par :

$$1 \leqslant i \leqslant n+1, \; x_i = \sum_{j=1}^{k+1} a_{ij} \lambda_j,$$

où les λ_j appartiennent à K et ne sont pas tous nuls. Ces $n+2$ formules définissent une bijection entre P,+,(E) et P(F) qui est une représentation paramétrique de la variété projective.

Dans le cas particulier de $\mathbf{P}_n(\mathbf{R})$, où $\mathbf{R}^{n+1}$ est muni de la base canonique, le plongement, indiqué **ci-dessus,** de l'espace affine de dimension n identifié à $\mathbf{R}^n$ dans $\mathbf{P}_n(\mathbf{R})$ fait correspondre au point $(\lambda_1$, A,, A,,) le point de coordonnées homogènes $(\lambda_1, \lambda_2, \ldots$ A,,, A,,,,) ; les « points à l'infini » de $\mathbf{P}_n(\mathbf{R})$ sont caractérisés par la condition $\lambda_{n+1} = 0$ et forment donc un hyperplan projectif.

La *géométrie projective* est l'étude des espaces projectifs et des variétés linéaires projectives, ainsi que des invariants par le groupe projectif.

JACQUES MEYER

PROJECTIVES APPLICATIONS

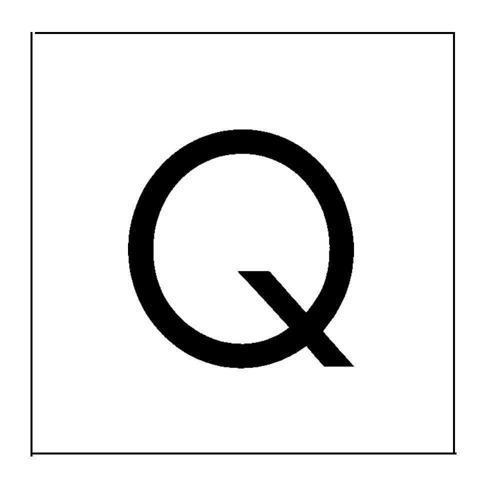
Soit E et F deux espaces vectoriels sur un même corps commutatif K, P(E) et P(F) les espaces projectifs déduits de E et de F, f une application linéaire de E dans F et N = ker (f) le noyau de f. Comme l'image par f d'une droite de E non contenue dans N est une droite de F, la restriction de f à E—N est compatible avec les relations d'équivalence sur E—N et F' = F—{0}. On peut donc déduire de f une application g de P(E) – P(N) dans P(F) par passage au quotient. L'application g est dite application linéaire projective ou

encore, par abus de langage, application projective de P(E) dans P(F).

Notons quefet λf, où A est un scalaire non nul, donnent la même application déduite. Réciproquement, si l'on se donne une variété linéaire projective P(N) et une application projective g de P(E)-P(N) dans P(F), toutes les applications linéaires dont g est déduite s'obtiennent à partir de l'une d'entre elles par multiplication par un scalaire non nul. Si l'on considère les applications projectives bijectives de P(E) dans P(F), on voit aisément que :

- les applications linéaires dont une application projective bijective est déduite sont elles-mêmes bijectives ;
- la composée de deux applications projectives bijectives est une application projective bijective ;
- les applications projectives bijectives de P(E) sur P(E) forment un groupe, appelé groupe projectif de P(E) et noté **PGL**(E) ; lorsque E = K^{n+1}, ce groupe se note $\mathbf{PGL}_n(K)$ ou $\mathbf{PGL}(n,K)$;
- lorsque les espaces projectifs P(E) et P(F) sont de dimension finie, et si leurs dimensions sont égales, toute application projective injective de P(E) dans P(F) est bijective et donc inversible. De plus, on a le théorème suivant : la donnée dans P(E) d'une famille $(\pi'(f_i)$ de $n+2$ points, formant un repère projectif, et dans P(F) d'une famille $(\pi'(f_i)$ de $n + 2$ points, formant un repère projectif, détermine une application projective et une seule de P(E) dans P(F), appliquant $\pi(e_i)$ sur $\pi'(f_i)$. De **plus,** cette application est bijective.

JACQUES MEYER



QUADRATIQUES FORMES

La notion de forme quadratique intervient dans toutes les parties des mathématiques. Elle est à la base de la géométrie euclidienne et de la mécanique classique (énergie cinétique), et aussi de la notion d'espace de Hilbert, de la théorie spectrale et de leurs nombreuses applications à l'analyse fonctionnelle (équations différentielles, aux dérivées partielles ou intégrales). Elle est étroitement liée au concept de dualité. Enfin, l'étude arithmétique des formes quadratiques a été le point de départ de la théorie des nombres algébriques et a eu d'importantes répercussions sur la théorie des fonctions automorphes.



1. Généralités

En algèbre classique, on appelle « forme n-aire de degré r » un polynôme *homogène* de degré r par rapport à n variables ; pour $r = 1$, on dit « forme linéaire » et, pour

$r = 2$, on dit « forme quadratique ». Dans la mathématique actuelle, on généralise la notion de forme quadratique comme on a généralisé celle de forme linéaire (cf. algèbre LINÉAIRE ET MULTILINÉAIRE) : étant donné un anneau commutatif A et un A-module M, on considère les applications Q de M dans A qui vérifient une relation de la forme :

$$(1)\quad Q(\lambda x + \mu y) = \lambda^2 A(x,y) + \lambda\mu B(x,y) + \mu^2 C(x,y),$$

quels que soient x et y dans M, λ et μ dans A. En donnant à λ et μ les valeurs 0 ou 1, on voit aussitôt que :

$$A(x,y) = Q(x),\ C(x,y) = Q(y)$$

et :

$$B(x,y) = Q(x+y) - Q(x) - Q(y);$$

en exprimant de plusieurs manières $Q(x + y + z)$, pour x, y et z dans M, on voit sans peine que l'expression :

$$D(x,y,z) = B(x+y,z) - B(x,z) - B(y,z)$$

est symétrique en x, y et z et que l'on a par suite :

$$D(\lambda x, \lambda y, \lambda z) = \lambda^3 D(x,y,z);$$

d'autre part, on a :

$$D(\lambda x, \lambda y, \lambda z) = \lambda^2 D(x,y,z),$$

donc $D(x, y, z) = 0$ lorsque A est sans diviseur de zéro et contient au moins trois éléments. Pour un anneau commutatif A quelconque, on dit que Q est une *forme quadratique* sur M si $D(x, y, z) = 0$ quels que soient x, y ct z dans M, c'est-à-dire si B est une *forme bilinéaire* (nécessairement symétrique). On dit que cette forme bilinéaire est *associée* à la forme quadratique Q. Si, dans l'anneau A, l'équation $2\xi = \alpha$ a une solution unique pour tout $\alpha \in A$, toute forme bilinéaire symétrique B sur M X M détermine inversement une forme quadratique Q à laquelle elle est associée, puisque :

$$B(x,x) = Q(2x) - 2\,Q(x) = 2\,Q(x).$$

La définition précédente montre par récurrence que, si $a_1, \dots, a_m$ sont des éléments de M et si $\xi_1, \dots, \xi_m$ sont des scalaires de A, on a :

$$(2)\quad Q\Big(\sum_{j=1}^{m} \xi_j a_j\Big) = \alpha_{11}\xi_1^2 + \dots + \alpha_{mm}\xi_m^2 + \sum_{i<j} \alpha_{ij}\xi_i\xi_j,$$

où $\alpha_{ij} = B(a_i, a_j)$ pour $i \neq j$ et $\alpha_{ii} = Q(a_i)$, donc $B(a_i, a_i) = 2\,\alpha_{ii}$. En particulier, si les a_i forment une hase de M, on retrouve la définition classique des formes quadratiques.

Exemples

Si l'on a été amené à donner une définition aussi générale, c'est parce que l'on rencontre naturellement des formes quadratiques de types très variés dans les applications. L'exemple le plus connu de forme quadratique est le « carré scalaire », dont l'étude est exactement la géométrie euclidienne. Deux des parties les plus importantes des mathématiques contemporaines, la géométrie riemannienne et la théorie des espaces de Hilbert, sont des extensions de cette étude dans deux directions : la forme quadratique est « infinitésimale » en géométrie riemannienne, et l'espace où elle est définie est de dimension infinie dans la théorie hilbertienne.

Dans tous ces cas, la forme est « positive non dégénérée » (cf. *infra*, chap. 2). Mais les formes non dégénérées non positives n'ont pas moins d'importance : leur théorie (pour les espaces de dimension

finie) a deux « traductions » classiques : la théorie des coniques, des quadriques et de leurs généralisations aux dimensions supérieures, et d'autre part les géométries « non euclidiennes » (cf. GROUPES - Groupes classiques et géométrie ; QUADRIQUES) ; l'aspect « infinitésimal » de cette théorie est la théorie des espaces pseudo-riemanniens, qui est à la base de la théorie de la relativité. Les formes quadratiques à coefficients complexes correspondent aux quadriques (et leurs généralisations) dans les espaces complexes ; et c'est une forme à coefficients complexes, la forme de Killing, sur laquelle repose la classification des groupes de Lie semi-simples.

L'étude des formes quadratiques à coefficients entiers, débutant avec Fermat et Euler, a été le ferment le plus actif dans le développement de la théorie des nombres : la théorie des formes binaires, équivalente à celle des corps quadratiques, a été, avec Gauss, le point de départ de la théorie des nombres algébriques ; celle des formes quaternaires est étroitement liée à la théorie arithmétique des quaternions et celle des formes à un nombre quelconque de variables est à l'origine du développement moderne de la théorie des groupes arithmétiques et des fonctions modulaires à n variables.

Tout récemment, les formes quadratiques ont reçu des applications plus inattendues. En topologie différentielle, c'est la considération d'une forme quadratique sur le corps à deux éléments $\mathbf{F}_2$ qui permet de définir un nouvel invariant, grâce auquel on a pu donner le premier exemple d'une variété topologique non susceptible d'être munie d'une structure différentielle (M. Kervaire) ; d'autres formes quadratiques interviennent en cohomologie (théorie de l'index) et en K-théorie, et l'on est même amené pour certaines questions à généraliser la notion de forme quadratique en considérant des applications de M dans un second A-module M' (« applications quadratiques »). L'application la plus imprévue est sans doute celle qui permet d'exclure *a priori* certains entiers N dans la recherche des plans projectifs finis (non desarguiens) ayant N + 1 points : on montre en effet que, s'il existe un tel plan, alors il y a une matrice carrée A d'ordre $N^2 + N + 1 = n$ à coefficients entiers telle que ${}^tA \cdot A = B$, où B est une matrice d'ordre n ayant tous ses éléments égaux à 1, sauf ceux de la diagonale principale égaux à N + 1 (Bruck-Ryser) : la théorie de Minkowski-Hasse (cf. *infra*, *Résultats spéciaux*, in chap. 2) donne des conditions arithmétiques de possibilité d'une telle relation qui permettent d'exclure certaines valeurs de N.

Transformation des formes quadratiques

La notion fondamentale à la base de toute la théorie des formes quadratiques est celle de *transformée* d'une telle forme par une application linéaire : si M et N sont deux A-modules, si $g : M \longrightarrow N$ est une application linéaire et Q une forme quadratique sur N, $x \mapsto Q(g(x))$ est une forme quadratique sur M, dite transformée de Q par g ; si B est la forme bilinéaire associée à Q, la forme :

$$(x, y) \mapsto B(g(x), g(y)),$$

associée à $Q \circ g$, est dite transformée de B par g. On se bornera dans toute la suite aux A-modules *de type fini* (pour la théorie hilbertienne, voir l'article théorie SPECTRALE). Lorsque M est un A-module *libre*, pour toute base (e,), $1 \leqslant j \leqslant n$, de M, la matrice carrée symétrique $T = (B(e_j, e_k))$ est appelée la matrice de B (ou de Q) par rapport à cette base ; si N est un second

A-module libre, si (f_i), $1 \leqslant i \leqslant m$, est une base de N et X la matrice de type (m, n) d'une application linéaire g de M dans N relativement aux bases choisies, alors la matrice de la transformée $Q \circ g$ de Q par rapport à (f_i) est ${}^tX.T.X$.

Les problèmes qui se posent naturellement dans la théorie des formes quadratiques sont les suivants.

A) Étant donné deux formes quadratiques Q_1 et Q_2 sur des A-modules M, et M_2, la forme Q_2 est-elle transformée de la forme Q_1 ? On dit encore alors que « Q_1 représente Q_2 ». En particulier, si M, et M_2 sont libres et si T_1 et T_2 sont les matrices de Q_1 et de Q_2 par rapport à des bases, une réponse positive à la question entraîne l'existence d'une matrice X à éléments dans A telle que :

$$(3) \qquad {}^tX.T_1.X = T_2;$$

inversement, cette existence entraîne que Q_2 est transformée de Q_1 lorsque, dans A, l'équation $2\xi = \alpha$ a toujours une solution unique. On notera que, dans ce dernier cas, lorsqu'on prend $M_2 = A$ de sorte que $T_2 = (a)$ est une matrice à un seul élément, résoudre l'équation *(3)* revient à trouver dans M, les solutions de $2Q_1(x) = \alpha$ que l'on appelle « représentations de l'élément $\alpha/2$ par Q_1 ».

B) « Classer » les formes quadratiques sur un module M pour diverses sortes de relations d'équivalence entre ces formes. De façon précise, on se donne un sous-groupe Γ du groupe des bijections linéaires de M sur lui-même, et on considère comme équivalentes deux formes quadratiques transformées l'une de l'autre par une application $g \in \Gamma$. On peut encore dire que l'ensemble des formes quadratiques sur M est un A-module $\mathfrak{Q}(M)$ dans lequel le groupe Γ ***opère*** *linéairement*, et on cherche les ***orbites*** de Γ pour cette action.

C) Étude du groupe de toutes les bijections linéaires de M qui transforment une forme quadratique en elle-même. Nous en avons donné d'importants exemples dans l'article GROUPES - Groupes classiques et géométrie. Notons simplement que les formes quadratiques sont exceptionnelles à cet égard parmi les formes de degré > 1 ; pour les formes de degré $\geqslant 3$, le groupe des transformations linéaires laissant invariant la forme est en général *fini.*

2. Formes quodratiques sur un corps

Nous distinguons deux cas, suivant que la caractéristique du corps de base K est distincte de 2 ou égale à 2.

Corps de caractéristique $\neq 2$

Résultats généraux

On peut se borner ici à considérer le problème de transformation d'une forme quadratique en une autre sous la forme (3). Un premier invariant est le *rang* de la matrice ***T*** d'une forme quadratique Q ; il est aussi appelé rang de Q ou rang de la forme bilinéaire associée B, et noté rg(Q) ou rg(B). C'est un entier qui peut prendre l'une quelconque des valeurs entre 0 et la dimension ***n*** de l'espace vectoriel V où est définie Q. On dit que la forme Q (ou B) est non *dégénérée* si rg (Q) $= n$; la forme bilinéaire B définit alors un isomorphisme φ de V sur son dual V* (cf. algèbre LINÉAIRE ET MULTILINÉAIRE) par la relation :

$$\langle x, \varphi(y)\rangle = B(x, y),$$

pour x et y dans V ; cela permet de définir dans V les notions de vecteurs orthogonaux, de sous-espace isotrope et de sous-espace totalement isotrope (cf. GROUPES - Groupes classiques et géométrie, chap. 3).

Un second invariant est *l'indice de Witt* $\nu \leqslant n/2$ (cf. GROUPES - Groupes classiques et géométrie, chap. 3) ; l'espace V se décompose en somme directe d'un sous-espace W de dimension $n - 2\nu$, ne contenant aucun vecteur isotrope $\neq 0$, et d'un sous-espace orthogonal à W, dans lequel l'indice de Witt de la restriction de Q à ce sous-espace est v ; pour n et v donnés, la classe d'équivalence de la forme Q_W, restriction de Q à W, détermine entièrement celle de Q, ce qui permet de ramener le problème d'équivalence au cas des formes *anisotropes* (c'est-à-dire d'indice 0).

On appelle *discriminant* de Q (ou de B) par rapport à une base de V le déterminant de la matrice de B par rapport à cette base ; comme la relation (3) entraîne :

$$\det(T_2) = \det(T_1)(\det(X))^2,$$

on voit que le discriminant dépend de la base choisie, mais sa classe *d(Q)* dans le groupe quotient K^*/K^{*2} du groupe multiplicatif K* de K par le sous-groupe des carrés dans K* est un invariant de Q.

Enfin, pour deux éléments a et β de K, on désigne par (a, β) l'algèbre de quaternions (généralisés), espace vectoriel de dimension 4 sur K ayant une base formée de 1 (élément unité) et de trois éléments x_1, x_2 et x_3 avec la table de multiplication :

$$x_1^2 = \alpha, \quad x_2^2 = \beta, \quad x_3^2 = -\alpha\beta,$$
$$x_1x_2 = -x_2x_1 = x_3, \quad x_1x_3 = -x_3x_1 = \alpha x_2,$$
$$x_2x_3 = -x_3x_2 = -\beta x_1.$$

C'est une algèbre simple de centre K, si $\alpha\beta \neq 0$. Cela étant, il y a toujours des bases (e_j), $1 \leqslant j \leqslant n$, de V orthogonales pour Q, autrement dit telles que :

$$(4) \quad Q(x) = \sum_{j=1}^{n} \alpha_j \xi_j^2, \; x = \sum_{j=1}^{n} \xi_j e_j \, ;$$

ainsi Q(x) est somme de « termes carrés ». On appelle *algèbre de Hasse* de Q pour cette base le produit tensoriel des algèbres de quaternions $(a,, a, \alpha_2 ,,, a,)$, pour $1 \leqslant j \leqslant n$, et on démontre que cette algèbre S(Q) ne dépend pas, à isomorphie près, de la base orthogonale choisie.

On peut prouver que, pour $n \leqslant 3$, les invariants rg(Q), *d(Q)* et *S(Q) caractérisent,* à équivalence près, les formes quadratiques sur un corps quelconque K (de caractéristique $\neq 2$) ; mais cela n'est plus exact pour $n \geqslant 4$. On n'a, dans ce cas. que des résultats pour des corps particuliers.

Résultats spéciaux

a) Le corps K est *algébriquement clos* ; un seul invariant suffit, le rang rg(Q) ; pour rg(Q) $= n$, on a $\nu = [n/2]$, partie entière de $n/2$.

b) Le corps K est le *corps* **R** *des nombres réels* ; pour toute base orthogonale (e_j) de V, si :

$$Q(x) = \sum_{j=1}^{n} \alpha_j \xi_j^2,$$

le nombre p (resp. q) des α_j qui sont > 0 (resp. < 0) est indépendant de la base choisie (« loi d'inertie » de Sylvester) ; on dit que (p, q) est la signature sig (Q) de Q ; les nombres p et q caractérisent les formes quadratiques à équivalence près ; on a

$\mathrm{rg}(Q) = p + q$; si $p + q = n$, on a v = inf (p, y) et $d(Q)$ est la classe de $(-1)^q$. Le groupe $\mathbf{R}^*/\mathbf{R}^{*2}$ a ici deux éléments.

c) Le corps K est *fini* ; dans ce cas, le groupe K^*/K^{*2} a encore deux éléments ; les invariants rg(Q) et $d(Q)$ caractérisent Q, à équivalence près ; si rg(Q) $= n$, on a v $\geqslant 1$ pour $n \geqslant 3$.

d) Le corps K est un corps *local* (cf. théorie des NOMBRES - Nombresp-adiques), d'idéal maximal $\mathfrak{P}$. Pour deux éléments α et β de K. le *symbole de Hilbert (a,* $\beta)_{\mathfrak{p}}$ est défini comme égal à 1 si l'équation $\alpha\xi^2 + \beta\eta^2 = 1$ a une solution dans K, comme égal à -1 dans le cas contraire (cf. DIVISIBILITÉ, chap. 4) ; et on montre que deux algèbres de quaternions (a, β) et (α', β') sont isomorphes si et seulement si on a $(\alpha, \beta)_{\mathfrak{p}} = (\alpha', \beta')_{\mathfrak{p}}$. On associe alors à la forme quadratique Q son *symbole de Hasse* S,(Q), égal par définition au produit des symboles de Hilbert $(\alpha_j, \alpha_1 \alpha_2 \dots \alpha_j)_{\mathfrak{p}}$ pour toute expression (4) de Q à l'aide d'une base orthogonale. On prouve que les invariants rg(Q), $d(Q)$ et S,(Q) caractérisent Q, à équivalence près. On a toujours v $\geqslant 0$ pour rg(Q) $\geqslant 5$.

e) Le corps K est un corps *de nombres algébriques* (cf. théorie des NOMBRES - Nombres algébriques). Pour toute place υ de K, le corps K se plonge canoniquement dans le corps local complété K_υ, et on peut donc considérer une forme quadratique Q sur K comme une forme quadratique Q_υ sur K,. La théorie est entièrement ramenée au cas des corps locaux par le *principe de Hasse* : pour qu'une forme quadratique Q' soit transformée d'une forme Q, il faut et il suffit que Q', soit transformée de Q_υ pour chaque place υ (finie ou à l'infini). Les invariants rg(Q), $d(Q)$, $S_{\mathfrak{p}}(Q_{\mathfrak{p}})$, pour toute place finie, et $\mathrm{sig}(Q_\upsilon)$ pour toute place réelle à l'infini. caractérisent donc Q, à équivalence près (théorème de Hasse-Minkowski). Les symboles $S_{\mathfrak{p}}(Q_{\mathfrak{p}})$ sont égaux à 1, sauf pour un nombre fini de places finies $\mathfrak{P}$, et on a la loi de réciprocité de Hilbert :

$$\prod_{\mathfrak{p}} S_{\mathfrak{p}}(Q_{\mathfrak{p}}) = 1.$$

Corps de caractéristique 2

Soit K un corps de caractéristique 2, V un espace vectoriel de dimension n sur K et Q une forme quadratique sur V. La forme bihnéaire B associée à Q est alors *alter&,* autrement dit $B(x, x) = 0$ pour tout $x \in V$; son rang rg (B) est par suite un nombre pair $2p$. Soit $V^\perp$ le sous-espace de V, formé des $x \in V$ tels que $B(x, y) = 0$ pour tout $y \in V$; sa dimension est $n \quad 2\,p$ et on a :

$$Q(\lambda x + \mu y) = \lambda^2 Q(x) + \mu^2 Q(y),$$

pour x et y dans V. L'ensemble V_0 **des** Y $\in V^\perp$ tels que Q(s) = 0 est donc un sous-espace vectoriel de V. Si $q \leqslant n \quad 2\,p$ est sa dimension, on dit que $2\,p + q$ est le *rang* rg (Q) et on appelle *défaut* de Q l'entier S(Q) = y = rg (Q) -- rg (B). Si U est un supplémentaire de V_0 dans $V^\perp$, si W est un supplémentaire de $V^\perp$ dans V et si l'on prend une base de V qui soit réunion d'une base symplectique (e_j), $1 \leqslant j \leqslant 2p$, de W (c'est-à-dire telle que $B(e_j, e_k) = 0$ sauf pour les couples (e_j, e_{j+p}), pour lesquels $B(e_j, e_{j+p}) = B(e_{j+p}, e_j) = 1$ pour $1 \leqslant j \leqslant p$). d'une base (e_j), $2p + 1 \leqslant$ j $\leqslant 2p + q$, de U et d'une base (e_j), $2p + q + 1 \leqslant$ j $\leqslant$ II, de V_0, alors on obtient, pour :

$$x = \sum_{j=1} \xi_j e_j,$$

l'expression de Q(x) suivante :

$$(5) \quad Q(x) = \sum_{j=1}^{p} (\alpha_j \xi_j^2 + \xi_j \xi_{j+p}^2 + \beta_j \xi_{j+p}^2) + \sum_{j=2p+1}^{2p+q} \gamma_j \xi_j^2,$$

où la relation :

$$\sum_{j=2p+1}^{2p+q} \gamma_j \xi_j^2 = 0$$

entraîne $\xi_j = 0$ pour $2p + 1 \leqslant j \leqslant 2p + q$. Le défaut q ne peut être > 0 que si K est *imparfait*, c'est-à-dire que le sous-corps K^2 de K formé par les carrés des éléments de K est distinct de K ; plus précisément, on a $q \leqslant [K : K^2]$.

Les entiers p et q sont évidemment des invariants de Q. On dit qu'un sous-espace L de V est *totalement singulier* si $Q(x) = 0$ dans L ; pour les formes de rang n, la dimension d'un tel espace est $\leqslant p$. Le maximum v des dimensions des espaces totalement singuliers est encore appelé *l'indice de Witt* de Q et est un invariant de cette forme.

Si, pour une base (e,) choisie comme plus haut, on forme l'élément :

$$\Delta(Q) = \sum_{j=1}^{p} \alpha_j \beta_j,$$

on dit que cet élément est le *pseudo-discriminant (ou invariant d'Arf)* de Q par rapport à cette base. Pour une autre base du même type, le pseudo-discriminant est de la forme :

$$\Delta(Q) + \xi^2 + \xi$$

pour un élément $\xi \in K$; comme les éléments $\xi^2 + \xi$ forment un sous-groupe P du groupe additif K, la classe $d(Q)$ de $\Delta(Q)$ dans le groupe quotient K/P est encore un invariant de Q.

Enfin, on peut généraliser aux corps de caractéristique 2 la notion d'algèbre de quaternions et obtenir ainsi pour Q un invariant qui généralise l'algèbre de Hasse définie *supra* (cf. *Résultats généraux*, in *Corps de caractéristique ≠ 2)*. Grâce à ces invariants, on peut, pour certains corps de caractéristique 2, obtenir une classification complète des formes quadratiques sur ces corps.

3. Réduction des formes quadratiques

Nous ne considérerons plus à partir de maintenant que des formes quadratiques non dégénérées sur le corps R des nombres réels, définies dans un espace $\mathbf{R}^r$, et nous nous intéresserons aux sous-groupes Γ du groupe linéaire $\mathbf{GL}(r, R)$ opérant à droite, par $(g, Q) \mapsto Q \circ g$, dans l'espace $\mathfrak{Q}(\mathbf{R}^r)$ de ces formes. Deux cas sont particulièrement étudiés, correspondant au groupe orthogonal $\Gamma = O(r, R)$ et au groupe $\Gamma = \mathbf{SL}(r, Z)$ des matrices inversibles de déterminant 1 à coefficients entiers. Nous renvoyons pour le premier cas à l'article théorie SPECTRALE, le problème étant celui de la réduction d'une forme quadratique (ou d'une « hyperquadrique ») à ses « axes ». La théorie de la « réduction » correspond au second cas. Comme les orbites de $\mathbf{GL}(r, R)$ dans $\mathfrak{Q}(\mathbf{R}^r)$ sont les ensembles de formes de signature donnée (p, q), avec $p + q = r$ (cf. *supra, Résultats spéciaux,* in *Corps de caractéristique ≠ 2*), il y a lieu de distinguer le cas des formes positives non dégénérées (c'est-à-dire $p = r$ et $q = 0$) et le cas des formes où p et q sont tous deux > 0 (dites aussi « indéfinies »).

Formes positives

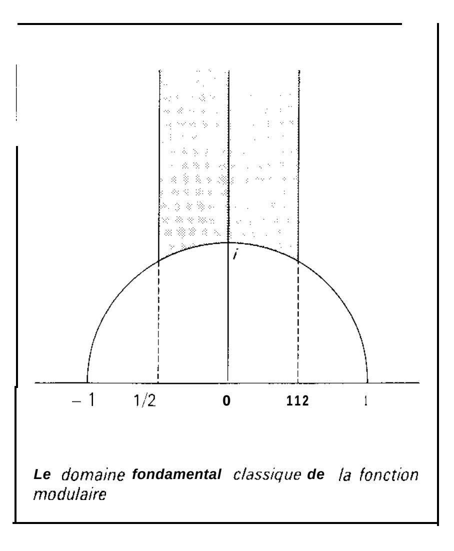
L'ensemble $\mathcal{H}$ (ou $\mathcal{H}_r$) de ces formes s'identifie à celui des matrices symétriques positives inversibles : c'est un « espace symétrique » $\mathcal{H}$ = K\G d'Élie Cartan, avec G = GL(r, R) et K = $O(r$, R) qui est le stabilisateur de la matrice unité. Le problème essentiel de la théorie de la réduction est de trouver dans $\mathcal{H}$ un « ensemble fondamental » $\mathfrak{S}'$ aussi « petit » que possible tel que toute orbite de Γ dans $\mathcal{H}$ le rencontre : il suffit de prendre l'image canonique dans $\mathcal{H}$ d'un ensemble $\mathfrak{S}$ C G tel que G = $\mathfrak{S} \cdot \Gamma$. Si l'on désigne par A le sous-groupe commutatif de G formé des matrices diagonales à termes $a_{ii} > 0$ et par N le sous-groupe des matrices triangulaires supérieures (n,) telles que $n_{ij} = 0$ si $j < i$ et $n_{ii} = 1$ pour $1 \leqslant i \leqslant r$, toute matrice $g \in$ G s'écrit d'une seule manière : $g = k \cdot a \cdot n$, avec $k \in$ K, $a \in$ A et $n \in$ N ; cette décomposition s'appelle « décomposition d'Iwasawa » (cf. GROUPES - Groupes de Lie, chap. 2).

On appelle *domaine de Siegel* $\mathfrak{S}_{t,u}$ dans G l'ensemble des matrices $k \cdot a \cdot n$, avec $a_{ii} \leqslant t \cdot a_{i+1,i+1}$ pour $1 \leqslant i \leqslant r-1$ et $n_{ij} \leqslant u$ pour $i < j$; une méthode remontant à Gauss (pour $r = 2$) et à Hermite prouve qu'il répond à la question posée, pour $t \geqslant 2/\sqrt{3}$ et $u \geqslant 1/2$. L'intérêt du choix d'un tel domaine fondamental $\mathfrak{S}$ est que son intersection avec SL(r, R) a une mesure de Haar *finie* ; d'autre part, si M est un ensemble de matrices m à coefficients entiers de déterminants *bornés*, alors l'ensemble $M_{\mathfrak{S}}$ des $m \in$ M telles que $\mathfrak{S} \cap \mathfrak{S} . m$ soit non vide est *fini* (Siegel). En outre, le fait que $\mathfrak{S}_{t,u}$, pour les valeurs de t et de u indiquées plus haut, soit un domaine fondamental entraîne l'inégalité d'Hermite

$$(6) \qquad \min_{x \in Z^r - \{0\}} Q(x) \leqslant (4/3)^{(r-1)/2} (\det Q)^{1/r},$$

où le déterminant est celui de la matrice de Q par rapport à la base canonique de R". Enfin, cela entraîne aussi que le groupe Γ est *de type fini.*

Un procédé de « réduction » plus fin, dû à Minkowski, fournit dans $\mathcal{H}$ un domaine fondamental plus petit que les domaines de Siegel $\mathfrak{S}'_{t,u}$, qui a la propriété de ne pouvoir avoir que des points frontières en commun avec ses transformés par Γ. Pour $r = 2$, en écrivant une forme quadratique positive $a(x + \tau y)(x + \bar{\tau} y)$, avec $a > 0$ et τ nombre complexe tel que Im $\tau > 0$, on identifie l'espace $\mathcal{H}^{(1)}$ des formes quadratiques positives, définies à un facteur constant près, au demi-plan Im $\tau > 0$; la réduction de Minkowski donne alors le domaine fondamental classique défini par $\tau \geqslant 1$ et Re $\tau \leqslant 1/2$ et représenté par la figure.



Le domaine fondamental classique de la fonction modulaire

Formes « indéfinies »

Si Q est une forme quadratique de signature (p, q) avec $p + q = r$ et $pq \neq 0$, on ne peut plus poser !e problème de !a « réduction » comme pour les formes positives. En effet, le groupe orthogonal O(Q),

sous-groupe de $GL(r, R)$ laissant Q invariante, est ici tel que $O(Q) \cap SL(n, Z)$ soit *infini*, et la propriété de finitude de Siegel ne peut donc être vérifiée pour aucun ensemble $\mathfrak{S}$ non vide. La notion de « réduction » qu'il faut introduire ici est une découverte célèbre d'Hermite. On ordonne l'ensemble $\mathcal{H}$ des formes quadratiques positives non dégénérées par la condition que $Q_1 \leqslant Q_2$ signifie que $Q_2 - Q_1 \in \mathcal{H}$; pour une forme quadratique indéfinie donnée Q, on appelle *majorante d'Hermite* de Q une forme $Q' \in \mathcal{H}$ telle que Q(s) $\leqslant$ Q'(x) pour tout $x \in \mathbf{R}^r$ qui est *minimale* dans l'ensemble des formes de $\mathcal{H}$ ayant cette propriété. Si $\mathcal{H}(Q)$ est l'ensemble des majorantes d'Hermite de Q, alors Se(Q) est encore un espace symétrique K C G, avec G = O(Q) et K sous-groupe compact maximal isomorphe à $O(p)$ X O(q). La forme Q est alors dite *réduite* au sens d'Hermite si l'intersection de X(Q) et d'un domaine de Siegel $\mathfrak{S}'$ dans $\mathcal{H}$ n'est pas vide, ou encore si Q est l'image par des opérateurs appartenant à un domaine de Siegel $\mathfrak{S}$ dans G de la forme canonique :

$$\xi_1^2 + \xi_2^2 + \dots + \xi_p^2 - \xi_{p+1}^2 - \dots - \xi_{p+q}^2 ;$$

le théorème de finitude fondamental est que, pour tout $\alpha > 0$ dans R, l'ensemble des formes réduites dont la matrice est de la forme αX, où X a ses éléments *entiers,* est un ensemble@.

Ces résultats ont été considérablement généralisés au cours de ces dernières années. On y remplace $GL(r, R)$ par le groupe des points réels $G_{\mathbf{R}}$, d'un groupe algébrique réductif G défini sur Q et Γ par un sous-groupe « arithmétique » de G : le problème fondamental est l'étude de l'espace homogène $G_{\mathbf{R}}/\Gamma$, et notamment l'obtention de critères pour que cet espace soit compact, ou de volume fini. ainsi que la preuve d'existence de « domaines fondamentaux » ayant des propriétés de finitude généralisant celles qui sont décrites ci-dessus (A. Borel-Harish-Chandra).

4. Formes quadratiques sur $\mathbf{Z}^n$

On se borne aux formes quadratiques sur $\mathbf{Z}^n$ non dégénérées, qui s'écrivent sous la forme Q : x $\mapsto B(x, x)$, où B est une forme bilinéaire sur $\mathbf{Z}^n$ X $\mathbf{Z}^n$ à valeurs dans Z ; la forme bilinéaire associée à Q est donc 2 B, et ce qu'on appelle la matrice de Q est ici la matrice de B (et non celle de 2 B) par rapport à la base canonique de $\mathbf{Z}^n$; c'est par suite une matrice symétrique non dégénérée arbitraire à coefficients entiers, Le problème fondamental est l'étude de l'équation (3), où T_1 et T_2 sont deux telles matrices, d'ordres respectifs n et $m \leqslant n$, et où la matrice inconnue X est une matrice de type (m, n) à coefficients entiers. Pour $m = n$, les matrices T_2 pour lesquelles (3) a une solution constituent la *classe* de T_1.

Une autre manière de présenter l'étude des formes quadratiques sur $\mathbf{Z}^n$ est de considérer une forme quadratique non dégénérée fixe sur $\mathbf{R}^n$. Si B est la forme bilinéaire symétrique associée, on considère les *réseaux* E dans $\mathbf{R}^n$, à savoir les Z-modules de type fini engendrant l'espace $\mathbf{R}^n$, tels que $B(x, y)$ soit entier pour x et y dans E ; deux tels réseaux sont isomorphes s'ils se déduisent l'un de l'autre par une transformation orthogonale (pour B). Comme tout réseau est un Z-module libre (donc isomorphe à $\mathbf{Z}^n$), les diverses bases de E correspondent aux formes quadratiques sur $\mathbf{Z}^n$ formant une *classe d'équivalence*. L'avantage de cette présentation est qu'elle s'étend au cas où l'on remplace Z par l'anneau des entiers d'un

corps de nombres algébriques ; les réseaux sur un tel anneau ne sont plus nécessairement des modules libres.

Dans l'étude des formes quadratiques sur $\mathbf{Z}^n$, on est amené à chercher à étendre le « principe de Hasse » de la théorie des formes quadratiques sur $\mathbf{Q}^n$. Les matrices X à coefficients entiers figurant dans l'équation (3) peuvent être considérées comme ayant leurs éléments dans l'un quelconque des anneaux d'entiers p-adiques $\mathbf{Z}_p$, ou dans R, et l'existence de solutions X à coefficients entiers implique donc celle de solutions X dans chacun de ces anneaux. Mais, ici, la réciproque n'est plus exacte ; les formes quadratiques $x^2 + 55y^2$ et $5x^2 + 11y^2$ sont équivalentes dans R et dans tous les $\mathbf{Z}_p$, mais non dans Z (la première représente 1. mais non la seconde). On est donc amené à envisager une notion d'équivalence moins stricte que celle qui est définie ci-dessus : deux matrices symétriques non dégénérées T_1 et T_2 correspondant à des formes quadratiques sur $\mathbf{Z}^n$ sont dites appartenir au même *genre* si l'équation (3) a, dans chaque $\mathbf{Z}_p$, une solution X_p (dépendant de p) ainsi qu'une solution dans R (ce qui signifie que les formes quadratiques correspondantes ont même indice). On déduit de la théorie de la réduction qu'un genre ne contient qu'un nombre *fini* de classes.

L'étude approfondie de l'équation (3) dans Z repose sur des méthodes analytiques, où la formule sommatoire de Poisson (cf. DISTRIBUTIONS, chap. 4) joue un rôle prépondérant. Il y a lieu de distinguer le cas des formes positives du cas des formes « indéfinies ».

Formes positives

Si S et T sont des matrices symétriques correspondant à des formes positives non dégénérées sur $\mathbf{Z}^n$, d'ordres respectifs n et m, avec $m \leqslant n$, on note N(S, T) le nombre de solutions en matrices X sur Z de l'équation ${}^tX \cdot S \cdot X = T$, nombre qui est *fini* et ne dépend que des classes de S et de T. On ne connaît pas de formule donnant ce nombre pour n et m quelconques, mais Siegel en a obtenu une expression « moyenne » qui fait intervenir non seulement la classe de S, mais toutes les classes du *genre* de S. Désignant par S, des représentants de ces classes, on pose :

$$\mu(S, T) = \frac{N(S, T)}{N(S, S)}, \tag{7}$$

et la formule de Siegel s'écrit :

$$\sum_j \mu(S_j, T) = \gamma(S) \prod_\nu \alpha_\nu(S, T), \tag{8}$$

où $\gamma(S)$ est la « masse » du genre de S au sens d'Eisenstein-Minkowski, c'est-à-dire le nombre :

$$\sum_j \frac{1}{N(S_j, S_j)},$$

somme des inverses des ordres des groupes d'automorphismes de S_j. Au second membre de (8), ν parcourt l'ensemble des « places » (finies ou non) de $\mathbf{Q}$, et $\alpha_\nu(S, T)$, qui ne dépend que des genres de S et de T, « mesure » en un certain sens l'ensemble des solutions de l'équation ${}^tX.S.X = T$ dans Qv. On a, d'autre part, une formule explicite (remontant à Minkowski) pour $\gamma(S)$:

$$\gamma(S) \prod_n \alpha_p(S, S) = 2\,\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{2}{2}\right)\dots\Gamma\left(\frac{n}{2}\right)}{\pi^{n(n+1)/4}} = (\det S)^{(n+1)/2}, \tag{9}$$

ce qui permet, en faisant $T = S$ dans (8), d'obtenir le nombre de classes dans le genre de S.

La formule (8) pour $T = S$ a une interprétation remarquable dans la théorie des

groupes « adéliques ». Si A est le groupe des adèles de Q (cf. théorie des NOMBRES - Nombres algébriques), on note G_Q, G_v et G_A les groupes des matrices carrées X à coefficients dans Q, Q_v et A respectivement vérifiant les relations $\det(X) = 1$ et ${}^tX.S.X =$ S. On définit un sous-groupe ouvert Go de G_A comme produit :

$$\prod G_\Omega^{(v)},$$

où v parcourt l'ensemble des places de Q, où $G_\Omega^{(v)} = G_v$ lorsque $v = \infty$ est la place à l'infini et où, pour chaque nombre premier p, l'ensemble $G_\Omega^{(p)}$ est l'ensemble des matrices de G_p à coefficients *entiers* yadiques.

On voit alors que les classes du genre de S correspondent biunivoquement aux ~ou-I&s *classes de* G_A suivant les sous-groupes G_Ω et G_Q : si U_j sont des représentants de ces doubles classes, de sorte que G_A est la réunion des $G_\Omega U_j G_Q$, on désigne par $R_p^{(j)}$, pour chaque nombre premier p, le transformé dans Q_p^n du réseau Z_p^n par l'automorphisme (U, $^1)_p$, projection de U_j^{-1} sur G,,. Il y a alors un réseau et un seul $R^{(j)}$ dans Q^n dont l'adhérence dans Q_p^n est $R_p^{(j)}$ pour tout p ; la matrice S_j est celle qui correspond à la forme quadratique lorsqu'on prend pour base de Q^n une base (sur ***Z)*** du réseau $R^{(j)}$.

On peut définir sur G_A une mesure de Haar privilégiée m (cf. analyse HARMONIQUE, chap. 4), dite *mesure de Tamagawa*, coïncidant dans G_Ω avec le produit de mesures de Haar m_v sur les G,.. Soit alors $G_0(U_j)$ le sous-groupe discret de G,. projection du groupe $(U_j^{-1}G_\Omega U_j) \cap G_Q$; des raisonnements élémentaires de la théorie de la mesure de Haar donnent la relation :

(10) $m\ (G_A/G_Q) =$

$$\Big(\prod_p m_p(G_\Omega^{(p)})\Big)\Big(\sum_j m_\infty(G_\infty/G_0(U_j))\Big),$$

où, par abus de langage. les mesures de Haar m et m_∞ sur des quotients de G_A ou de G_∞ par des groupes discrets sont celles qui sont déduites canoniquement des mesures notées m et m_∞ sur G_A et G,. On constate alors que cette formule devient identique à la formule de Siegel (8) pour T = S, une fois que l'on a prouvé que le « nombre de Tamagawa » $m(G_A/G_Q)$ est *égal à* 2 ; la preuve de ce fait (qui peut se faire indépendamment des résultats de Siegel) nécessite le même genre de méthodes analytiques. On peut aussi obtenir de cette manière la formule générale (8) pour $n \geqslant 4$ et $m \leqslant n - 3$. En outre, cette méthode d'« adélisation » peut être considérablemcnt généralisée en remplaçant G par un groupe algébrique semi-simple défini sur Q et en considérant des sous-groupcs « arithmétiques » convenables de G (Tamagawa, A. Weil, T. Ono).

Formes indéfinies

Les développements précédents subsistent sans modification lorsqu'on remplace G par le groupe analogue correspondant à la matrice S d'une forme ***indéfinie à coeffi***cients entiers ; mais, comme ici les nombres N(S, T) sont infinis, il n'est plus possible d'interpréter la formule (10) et l'analogue pour $T \neq$ S de la même manière que pour les formes positives ; Siegel a montré comment le faire en interprétant, dans la formule (8), les nombres $\mu(S_j, T)$ comme des volumes de domaines fondamentaux pour certains groupes discontinus ou comme des limites de rapports de nombres de solutions comme dans (7), où l'on impose aux solutions d'être dans un domaine *borné* de Z^{mn} et où l'on fait ensuite tendre ce domaine vers l'espace tout entier.

Donnons un exemple de ce genre d'interprétation qui précise le théorème de

Meyer affirmant qu'une forme quadratique indéfinie à coefficients entiers et à cinq variables au moins a toujours des solutions non triviales. Si :

$$Q(x) = \sum_{k=1}^{n} a_k \xi_k^2, \quad n \geqslant 5$$

est une forme indéfinie à coefficients entiers, si :

$$Q_0(x) = \sum_{k=1}^{n} |a_k| \xi_k^2$$

est une majorante d'Hermite de cette forme et si on pose :

$$A(e) = \sum_{x} e^{-\pi\varepsilon Q_0(x)}, \varepsilon > 0$$

x parcourant l'ensemble infini des solutions de $Q(x) = 0$, alors cette série est convergente et, lorsque ε tend vers 0, le nombre $A(\varepsilon)$ croît indéfiniment et est équivalent à $C\varepsilon^{(n \; 1)/2}$, où C est une constante.

À d'autres égards, les formes indéfinies ont une théorie plus simple que les formes positives : le nombre des classes d'un genre est toujours une puissance de 2 (qui peut être arbitrairement grande), et les nombres $\mu_j(S, T)$ pour les classes d'un même genre sont tous *égaux*, ce qui donne leur valeur en vertu de la formule de Siegel lorsqu'on connaît le nombre de classes du genre. Dans certains cas, on a même une classification complète des réseaux correspondant aux formes quadratiques indéfinies : il en est ainsi pour les formes sur $\mathbf{Z}^n$ de déterminant ± 1. On les classe en deux types suivant que la forme quadratique ne prend que des valeurs paires (type 2) ou prend aussi des valeurs impaires (type 1). Les réseaux de type 1 sont isomorphes à $pI_+ \oplus qI_-$, où I_+ (resp. I_-) correspond à la forme quadratique x^2 (resp. x^2) sur Z et où p et q sont des entiers $\geqslant 1$; les réseaux de type 2 sont isomorphes à $\pm(pU \oplus q\Gamma_8)$, avec p et q entiers $\geqslant 0$, où U correspond à la forme quadratique $2\,x_1x_2$ sur $\mathbf{Z}^2$. Pour $n = 4k$, Γ_n est le réseau dans $\mathbf{Q}^n$ formé des (x_j), $1 \leqslant j \leqslant n$, tels que $2\,x_j$ et $x_i - x_j$ soient entiers pour tous les indices, et que l'entier :

$$\sum_{j=1}^{n} x_j$$

soit pair, la forme bilinéaire fixe prise sur $\mathbf{Q}^n$ étant :

$$\sum_{j=1}^{n} x_j y_j \, ;$$

on vérifie que la forme quadratique positive sur $\mathbf{Z}^n$ définie par Γ_{4k} est à coefficients entiers et ne prend que des valeurs paires si k est pair.

5. Formes quadratiques et fonctions modulaires

Lorsqu'on fait $m = 1$ dans la formule de Siegel (8), de sorte que ***Test*** réduite à un seul entier N, on obtient une « valeur moyenne » du nombre de solutions de l'équation Q(s) $=$ N dans $\mathbf{Z}^n$ pour une forme positive Q sur $\mathbf{Z}^n$; si l'on sait que le genre de S ne contient qu'une *seule* ***classe,*** ou si les nombres N(S,, T) sont *les mêmes* pour toutes les classes du genre de S, la formule (8) donne le nombre de solutions de $Q(x) = N$ pour tout N. Par exemple. si :

$$Q(x) = \sum_{j=1}^{n} x_j^2,$$

on sait depuis Eisenstein que ie genre de S n'a qu'une seule classe pour $n \leqslant 8$, mais ce n'est plus exact pour **17** $\geqslant 9$; pour $n = 16$,

il y a deux classes dans le genre, mais elles donnent la même valeur à $N(S_j, T)$. On déduit donc de la formule de Siegel (8) des expressions exactes pour le nombre de représentations de N comme somme de n carrés pour $n \leqslant 8$ ou $n = 16$.

La théorie des fonctions thêta et des formes modulaires donne des expressions remarquables pour le second membre de (8) pour $m = 1$. Soit Q(s) une forme quadratique positive non dégénérée sur $\mathbf{Z}^n$ et soit S sa matrice ; la série :

$$(11) \qquad \theta_s(z) = \sum_x \exp(\pi i \, Q(x) z),$$

où x parcourt $\mathbf{Z}^n$, est absolument convergente pour Im z > 0 et est donc une fonction holomorphe de z dans ce demi-plan. La formule sommatoire de Poisson permet de prouver l'identité de Jacobi générale :

$$(12) \ \theta_s(z) = ((-iz)^n \det S)^{1/2} \theta_{s-1}\left(-\frac{1}{z}\right)$$

Bornons-nous, pour simplifier, au cas où $\det(S) = 1$ et où les termes diagonaux de S sont pairs ; on montre que cela implique que n est un multiple de 8 ; la relation (12) s'écrit alors :

$$\theta_s\left(-\frac{1}{z}\right) = z^{n/2} \theta_s(z),$$

et d'autre part :

$$\theta_s(z + 1) = \theta_s(z).$$

Or, dans le demi-plan supérieur Im $z > 0$, les transformations $z \mapsto z + 1$ et $z \mapsto 1/z$ engendrent le *groupe modulaire* de toutes les transformations :

$$z \mapsto \frac{az + b}{cz + d},$$

avec *a, b, c, dentiers* et $ud - hc = 1$; c'est un groupe discontinu dont un domaine fondamental est donné par la figure. Une *forme modulaire* de poids 2 *k,* pour *k* entier, est une fonctionfholomorphe dans Im z > 0 telle que :

$$f(z) = (cz + d)^{-2k} f\left(\frac{az + b}{cz + d}\right),$$

pour toute transformation du groupe modulaire ; en outre, on impose à f la condition d'être holomorphe à l'infini, ce qui implique qu'elle est développable en série :

$$f(z) = \sum_{N=0}^{\infty} a_N e^{2\pi i N z}$$

convergente pour Im $z > 0$; on dit de plus que la forme est *parabolique* si $a_0 = 0$. La fonction θ_s est donc une forme modulaire de poids $n/2$.

Si $k > 1$, la *série d'Eisenstein :*

$$(13) \qquad G_k(z) = \sum_{(c,d)} (cz + d)^{-2k},$$

où (c, d) parcourt $\mathbf{Z}^2 - \{0\}$, est convergente (cas particulier d'une série de Poincaré pour le groupe modulaire) et est une forme modulaire de poids 2 *k,* telle que $G_k(\infty) = 2\,\zeta(2k)$; en posant $q = e^{2\pi i z}$, on montre que :

$$G_k(z) = 2\,\zeta(2k)\, E_k(q),$$

avec :

$$(14) \ E_k(q) = 1 + (-1)^k \frac{4k}{B_k} \sum_{N=1}^{\infty} \sigma_{2k-1}(N) q^N,$$

où B_k est le k-ième nombre de Bernoulli et $\sigma_{2k-1}(N)$ la somme des puissances (2 *k*- 1)-ièmes des diviseurs de N (cf. calculs ASYMPTOTIQUES, chap. 2).

On montre alors que l'on a $\theta_s = E_k + f_s$ où $k = n/4$ et où f_s est une

forme modulaire parabolique. Comme l'on peut écrire :

$$(15) \qquad \theta_s(q) = \sum_{N=0}^{\infty} r_s(N) q^N,$$

où v,(N) est le nombre de solutions de l'équation Q(s) = 2 N dans Z^n, on déduit de (14) l'expression asymptotique :

$$(16) \qquad r_s(N) = \frac{4k}{B_k} \sigma_{2k-1}(N) + O(N^k).$$

Pour $n = 8$ ou $n = 16$, on a $f_s = 0$; pour n = 24, on montre que $f_s = c_s F$, où F est la forme parabolique :

$$F(q) = q \prod_{N=1}^{\infty} (1 - q^N)^{24},$$

étroitement liée à la théorie des fonctions elliptiques, et c_s une constante dépendant de la classe de S et qu'on détermine en évaluant le coefficient $r_s(1)$; on montre, pour n = 24. qu'il y a vingt-quatre classes de formes quadratiques vérifiant les conditions indiquées ci-dessus. Siegel a montré que, pour $m = 1$, la matrice T étant réduite à l'entier N, la série génératrice dont le coefficient de q^N est le premier membre de la formule (8) s'exprime encore à l'aide de séries d'Eisenstein. 11 a ensuite étendu ce fait au cas où m est quelconque, en introduisant des « fonctions modulaires d'ordre m », où le demi-plan Im $z > 0$ est remplacé par le « demi-espace de Siegel » formé des matrices symétriques complexes d'ordre m, dont les parties imaginaires sont des matrices positives non dégénérées ; le groupe modulaire est remplacé par le groupe de transformations :

$$Z \mapsto (AZ + B)(CZ + D)^{-1}$$

du demi-espace de Siegel, la matrice :

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

d'ordre 2 m parcourant le groupe symplectique Sp (2 m, Z) sur Z ; la notion de série d'Eisenstein se généralise et permet d'étendre les formules précédentes.

JEAN DIEUDONNÉ

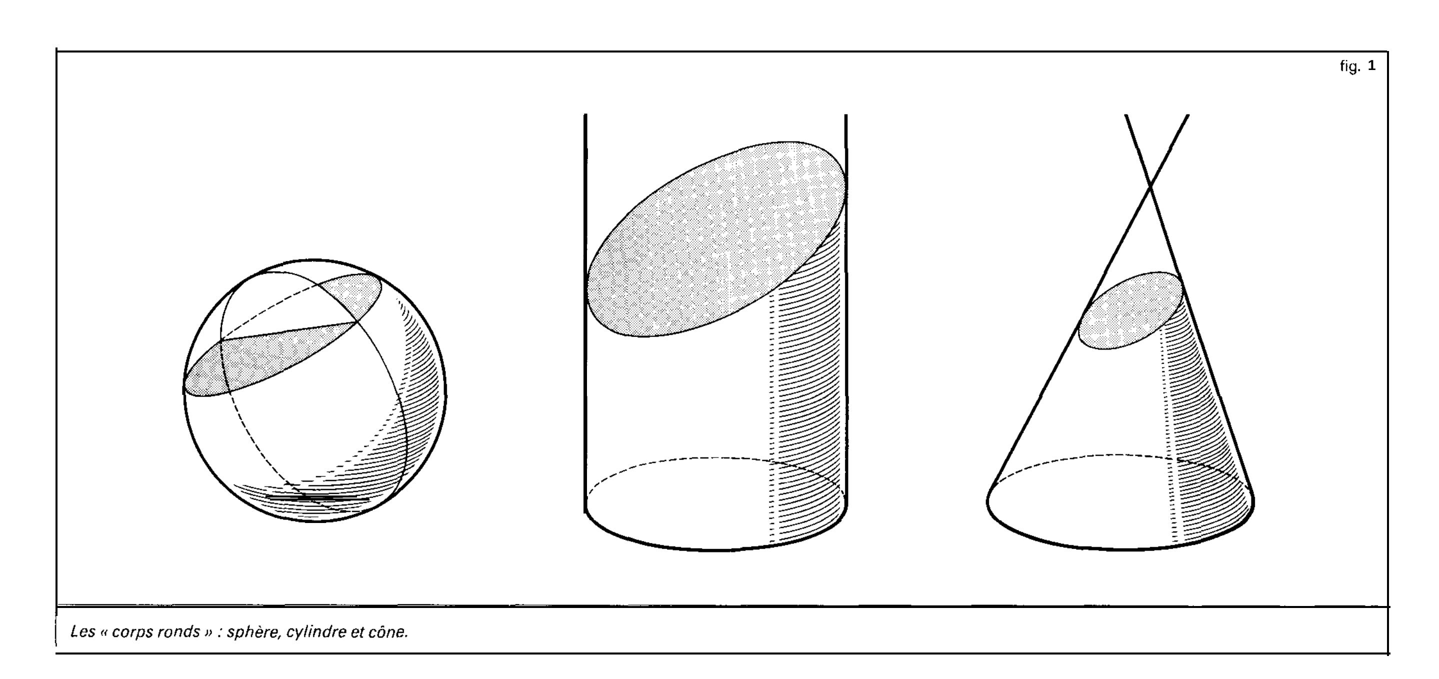
Bibliographie

A. BOREL, *Introduction aux groupes arithmétiques*, Hermann, Paris, 1969 / J. W. CASSELS, *Rational Quadratic Forms*, Acad. Press, New York, 1979 / M. KNESER, « Klassenzahlen definiter quadratischer Formen », in *Archiv der Mathematik*, n° 8, 1957 / 0. T. O'MEARA, *Introduction to Quadratic Forms*, Springer Verlag, New York-Berlin, 3' éd., 1973 / J.-P. SERRE, *Cours d'arithmétique*, P.U.F., Paris, 1970 / C. L. SIEGEL, *Gesammelte Abhandlungen*, 3 vol.. Berlin. 1966 / A. WEIL. *Sur lu théorie des formes quadratiques*, Bruxelles, 1962 / « Sur la formule de Siegel dans la théorie des groupes classiques », in *Acta mathematica*, n° 113, 1965 / E. WITT, « Theorie der quadratischen Formen in beliebigen Körpern », in *Journal de Crelle*, n° 176, 1937.

QUADRIQUES

Les surfaces de l'espace matériel, que nous connaissons par leur emploi, en architecture par exemple, étaient autrefois classées en « corps ronds » et « corps droits ». La sphère et le cube sont des surfaces typiques de ces deux familles.

Les corps ronds sont, essentiellement, la *sphère* déjà citée, le *cylindre* et le *cône* usuels (fig. 1). Étudiées individuellement, ces surfaces semblent n'avoir que peu de points communs : l'une est bornée, les deux autres ne le sont pas. Le cône possède un point remarquable (son sommet), alors que ie cylindre est totaiement homogène. Il est toutefois bien connu que les intersections de ces trois surfaces par des plans



Les « corps ronds » : sphère, cylindre et cône.

sont toujours des ***coniques,*** éventuellement dégénérées en couples de droites. L'adjectif conique, c'est-à-dire dessiné sur un cône, est à l'origine du nom donné à ces courbes (cf. **CONIQUES).**

Les propriétés très remarquables des coniques, qui constituent l'ensemble le plus riche de courbes simples, avaient conduit les Grecs à unifier partiellement les définitions et les démonstrations propres à chacune d'elles (ellipse, parabole et hyperbole). Seule la géométrie analytique cartésienne, pourtant, a permis de donner des coniques la définition essentielle : ce sont les *courbes* ***algébriques du second ordre,*** c'est-à-dire les ensembles de points dont les coordonnées (x, y) satisfont à une égalité de la forme :

$$P(x, y) = 0,$$

où P est un polynôme non nul du second degré. Suivant la nature des nombres x et y (qui sont réels ou complexes), on définit plusieurs types de coniques.

La généralisation de cette notion à l'espace de dimension trois est alors évidente. Les surfaces ainsi définies sont appelées *quadriques*. Leurs sections planes sont des coniques ; et cela les caractérise évidemment parmi les surfaces algébriques.



1. Cadre naturel de la théorie

Extensions diverses

Une quadrique est un ensemble de points satisfaisant à une égalité de la forme suivante où l'un au moins des six premiers coefficients n'est pas nul :

$$ax^2 + a'y^2 + a''z^2 + 2byz + 2b'xz + 2b''xy + 2cx + 2c'y + 2c''z + d = 0;$$

par exemple, une sphère (dont le centre a pour coordonnées u, v et w) a une équation du type :

$$x^2 + y^2 + z^2 - 2ux - 2vy - 2wz + (u^2 + v^2 + w^2 - r^2) = 0.$$

Certaines complications dans la théorie de ces surfaces conduisirent les mathématiciens du XIXe siècle à étendre. dans deux directions différentes. la notion de quadrique et des autres surfaces algébriques. Non seulement ils firent un usage systématique des coordonnées complexes, enrichissant ainsi considérablement le modèle mathématique issu des corps de l'espace matériel, mais ils élargirent le concept de point en considérant des « éléments à l'infini » par l'introduction d'une quatrième coordonnée t, nulle pour les nouveaux points.

Pour ces géomètres, une quadrique était donc finalement déterminée par une équation de la forme :

$$ax^2 + a'y^2 + a''z^2 + 2byz + 2b'xz + 2b''xy + 2cxt + 2c'yt + 2c''zt + dt^2 = 0.$$

Les « points » de ces quadriques sont, en réalité, non des points d'un espace classique (dit *affine*, de dimension trois sur le corps des nombres réels), mais des droites d'un espace vectoriel de dimension *quatre* sur le corps des nombres complexes. L'ensemble de ces droites, c'est-à-dire des sous-espaces vectoriels de dimension 1, est appelé le *projectifié complexifié* de l'espace usuel.

Cette extension donna une grande unité aux théorèmes : ainsi, une droite arbitraire coupe toujours une quadrique en un point au moins. Dans ce cadre, une théorie fort élégante de la polarité, directement généralisée à partir de considérations analogues sur les coniques, conduit à un grand nombre de propriétés (sur les plans tangents, par exemple), pour lesquelles on n'est plus tenu de distinguer un grand nombre de cas.

Quadriques et formes quadratiques

Ce passage d'un espace affine à un espace projectif - au prix d'une augmentation de la dimension - permit surtout de placer la théorie des quadriques dans son véritable cadre : celui des *formes quadratiques*. Conformément à une tendance actuelle, l'étude détaillée des coniques et des quadriques est aujourd'hui délaissée au profit de la notion plus générale d'*hyperquadrique*, définie comme un ensemble de droites vectorielles d'un espace E sur un corps algébriquement clos de caractéristique différente de deux ; une droite appartient à cet ensemble si et seulement

si les vecteurs qui la composent annulent une forme quadratique non nulle q. L'égalité :

$$q(Y) = 0$$

est une équation de l'hyperquadrique.

Cette généralisation est parfaitement typique de l'algébrisation d'une théorie géométrique ; en quelques pages, toutes les notions de conjugaison et leurs cas particuliers (hyperplans tangents ou symétries centrales, par exemple) peuvent être étudiés comme applications de la théorie de la conjugaison de vecteurs par rapport à la forme q.

Donnons un exemple : pour que deux points donnés, associés à deux vecteurs $\vec{v}$ et $\vec{v}'$ de l'espace E, forment une division harmonique avec deux points de l'hyperquadrique, il est nécessaire – et suffisant en général – que l'on ait l'égalité :

$$q(\vec{v} + \vec{v}') = q(\vec{v}) + q(\vec{v}').$$

Dans le cas particulier évidemment très utile – où l'espace E est de dimension finie, le calcul matriciel permet une traduction souple des calculs analytiques. C'est ainsi que l'équation d'une quadrique classique peut s'écrire :

$${}^tX . A . X = 0,$$

où X est la matrice colonne composée des quatre nombres x, y, z et t ; ${}^tX$ est la matrice ligne transposée de la précédente, et A une matrice symétrique réelle :

$$A = \begin{pmatrix} a & b'' & b' & c \\ b'' & a' & b & c' \\ b' & b & a'' & c'' \\ c & c' & c'' & d \end{pmatrix}$$

Une quadrique est dite *propre* si la forme quadratique associée q est non dégénérée, ce qui se traduit par le fait que le déterminant A = det A de la matrice A est non nul. Si A = 0, la quadratique est dite *impropre*.

2. Quadriques impropres

Il existe onze types différents de quadriques impropres, parmi lesquels on distingue trois familles principales : les cônes, les cylindres et les quadriques décomposées.

Cônes

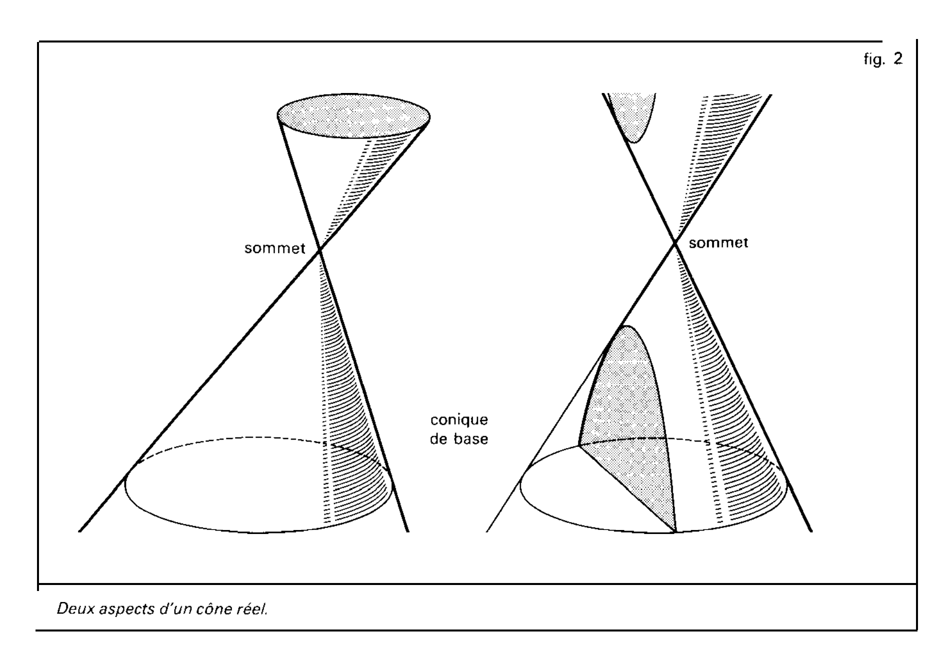
Les cônes sont obtenus à partir d'un sommet et d'une base, conique non décomposée dont le plan ne contient pas le sommet. Le cône de révolution est l'un d'eux ; on peut l'engendrer par rotation d'une droite autour d'une droite fixe qu'elle rencontre : un cône est constitué de deux nappes, c'est-à-dire de deux parties symétriques limitant des volumes convexes et reliées entre elles par le sommet commun (fig. 2).

Le sommet d'un cône est toujours réel. C'est le seul point de cette espèce si la conique de base est totalement non réelle ; le cône est alors dit imaginaire. Sinon, le cône est réel, et la nature de la conique de départ est sans importance.

Cylindres

Les *cylindres* sont des cônes dont le sommet est « à l'infini » : ils sont donc obtenus par des droites (génératrices) ayant une direction donnée qui rencontrent une conique non décomposée dont le plan n'est pas parallèle à la direction. Là aussi, le cylindre de révolution (qui sert de base à tant d'éléments architecturaux) en est l'exemple le plus simple.

Suivant que la conique est non réelle, ou une ellipse, ou une hyperbole, ou une parabole, le cylindre est dit imaginaire,



Deux aspects d'un cône réel.

elliptique, hyperbolique ou parabolique (fig. 3). Le cylindre de révolution est un cas particulier de cylindre elliptique. Les sections planes sont, en général, des coniques de même genre que la conique de base.

Quadriques décomposées

Les couples *de plans* sont des quadriques à la fois impropres et *décomposées* ; ce dernier qualificatif signifie simplement qu'il s'agit alors de la réunion de surfaces algébriques d'ordre inférieur à celui de la quadrique : ce sont donc des plans. Il en existe cinq sortes, selon que les deux plans sont réels et sécants, réels et parallèles, non totalement réels (mais transformés l'un en l'autre par une conjugaison des coordonnées) et sécants, non réels et parallèles, ou réels et confondus.

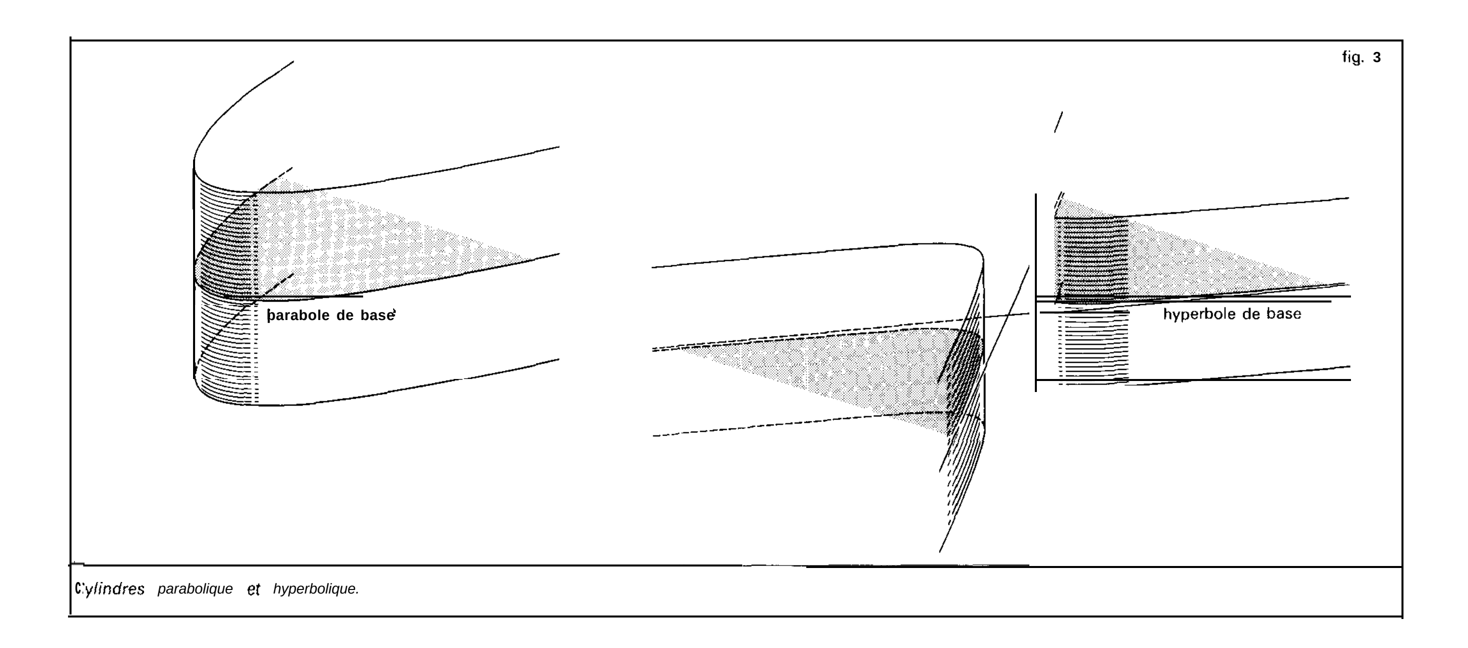
Pour qu'une quadrique soit décomposée, il est nécessaire et suffisant que l'on puisse écrire l'égalité :

$$\begin{vmatrix} a & b'' & b' \\ b'' & a' & b \\ b' & b & a'' \end{vmatrix} + \begin{vmatrix} a & b' & c \\ b' & a'' & c'' \\ c & c'' & d \end{vmatrix} + \begin{vmatrix} a & b'' & c \\ b'' & a' & c' \\ c & c' & d \end{vmatrix} + \begin{vmatrix} a' & b & c' \\ b & a'' & c'' \\ c' & c'' & d \end{vmatrix} = 0,$$

et que la quadrique soit impropre. Pour qu'elle soit décomposée en un plan double, il s'introduit une condition supplémentaire : l'ensemble des deux premières conditions est équivalent au fait que 0 doit être une ***valeur propre*** au moins double de la matrice A ; la troisième condition entraîne que 0 est alors une valeur propre au moins triple.

3. Quadriques propres

Les quadriques propres présentent moins de variété. Elles se classent également en trois familles (ellipsoïdes, hyperboloïdes et paraboloïdes) ayant chacune deux sous-familles.



Cylindres *parabolique* et *hyperbolique.*

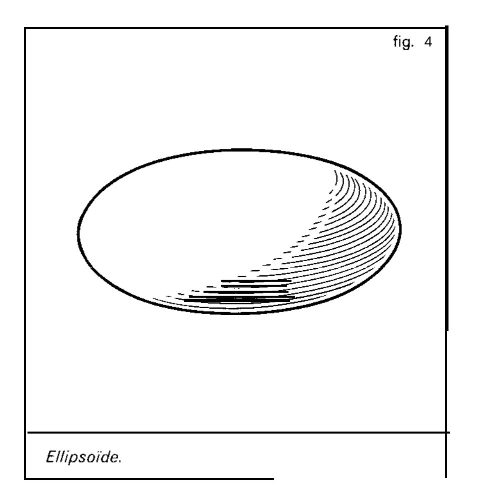
Ellipsoïdes

Les ellipsoïdes (fig. 4), par un changement d'axes approprié, peuvent se mettre sous la forme :

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=\pm t^2,$$

où a, b et c sont des nombres réels et strictement positifs. Le signe moins correspond à un ellipsoïde imaginaire, dont aucun point n'a toutes ses coordonnées réelles. Le signe plus correspond à l'ellipsoïde classique dont la partie réelle équivaut, grosso modo, à une sphère, surface en laquelle on peut le transformer par des affinités appropriées (de la même façon que l'on transforme une ellipse réelle en un cercle). Il est de révolution si deux des nombres a. b et c (b et c, par exemple) sont égaux; on le qualifie de sphérique si $a=b=c$, d'aplati si $a<b=c$, d'allongé si $a>b=c$.

Les sections planes d'un ellipsoïde sont en général des ellipses, réelles ou non.


fig. 4

Ellipsoïde.

Hyperboloïdes

Les hyperboloïdes ont une équation que l'on peut mettre sous l'une des deux formes suivantes :

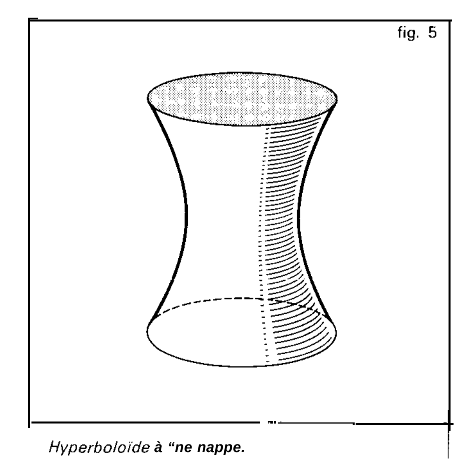
$$(H_1) \qquad \frac{x^2}{a^2} + \frac{y^2}{b^2} + t^2 = \frac{z^2}{c^2},$$

$$(H^2) \qquad \frac{x^2}{a^2} + \frac{y^2}{b^2} = t^2 + \frac{z^2}{c^2}.$$

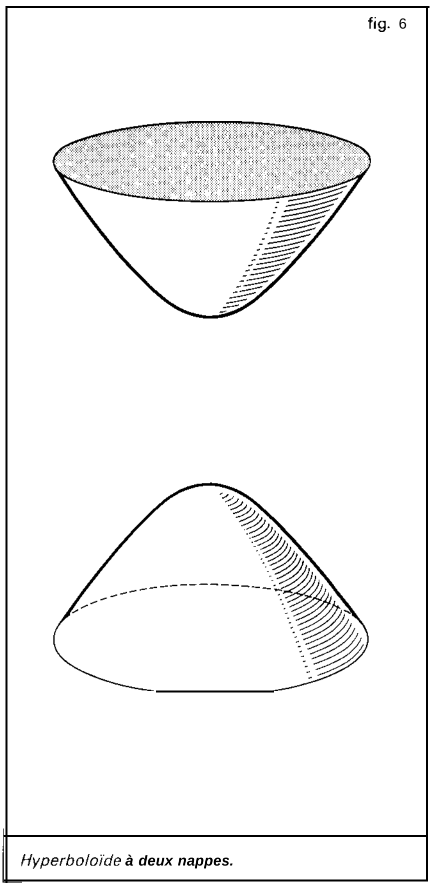
Le premier cas est celui de l'*hyperboloïde à une nappe* (fig. 5), qui est une surface connexe évoquant la forme d'une bobine. On peut le considérer, de deux façons différentes. comme réunion d'une famille de droites. les génératrices. Une affinité convenable, qui revient a égaler les coefficients a et b, le transforme en hyperboloïde de révolution, engendré par la rotation d'une droite autour d'une droite qu'elle ne rencontre pas.

Les sections planes sont des coniques de toutes espèces ; un plan tangent coupe l'hyperboloïde suivant deux droites sécantes, qui le séparent en deux parties situées de chaque côté de ce plan.

Le second cas est celui de l'*hyperboloïde à deux nappes* (fig. 6). qui admet deux


fig. 5

Hyperboloïde **à "ne nappe.**


fig. 6

Hyperboloïde **à deux nappes.**

nappes disjointes, connexes, limitant deux volumes convexes. Les génératrices d'une telle surface ne sont pas réelles, sauf éventuellement en leur point commun.

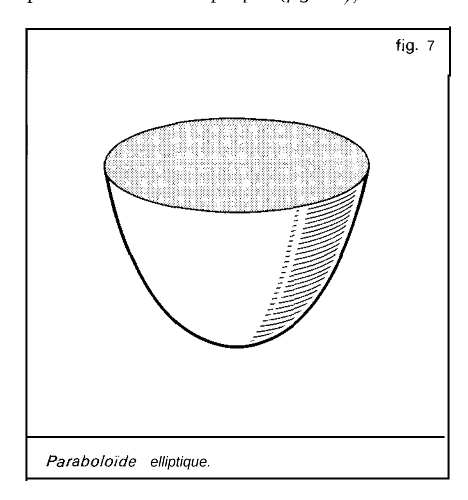
Paraboloïdes

Les paraboloïdes ont une équation que l'on peut mettre sous la forme :

$$\frac{x^2}{p} + \frac{y^2}{q} = 2zt,$$

où p et q sont deux nombres réels non nuls.

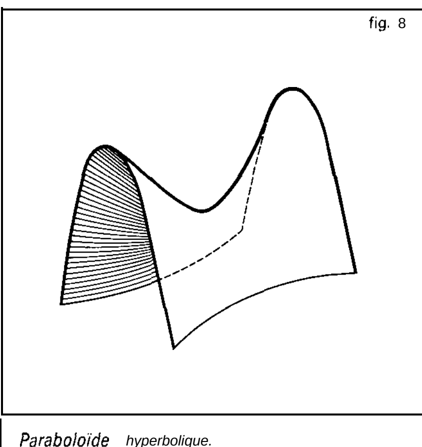
Si p et q sont de même signe, le paraboloïde est *elliptique (fig. 7)*, de révolution si $p = q$. Une affinité convenable peut toujours mettre le paraboloïde sous cette forme ; la surface résulte alors de la rotation d'une parabole autour de son axe.


fig. 7

Paraboloïde elliptique.

Les sections planes sont des paraboles ou des ellipses. Dans le cas d'un paraboloïde de révolution, une section plane se projette sur un plan orthogonal à l'axe suivant une droite ou un cercle.

Si p et q sont de signes différents, le paraboioïde est *hyperbolique (fig. 8).* C'est une surface assez remarquable, dont la forme évoque celle d'une selle de cheval.


fig. 8

Paraboloïde hyperbolique.

Une quadrique propre possède, comme nous l'avons vu à propos de l'hyperboloïde à une nappe, un double système de génératrices, Dans le cas du paraboloïde hyperbolique, les génératrices passant par un point réel sont réelles ; elles séparent la surface en deux parties situées de part et d'autre du plan tangent. Les sections planes sont des paraboles ou des hyperboles.

Le paraboloïde hyperbolique possède de nombreuses définitions géométriques très simples. Citons-en deux :

Si l'on se donne trois droites soumises à la **seule** condition d'être parallèles à un même plan, une droite variable qui rencontre ces trois droites engendre un paraboloïde hyperbolique ;

- Si l'on se donne deux droites quelconques, une droite variable qui les rencontre toutes les deux et reste parallèle à un plan donné engendre un paraboloïde hyperbolique.

Les plans apparaissant dans l'une ou l'autre de ces définitions sont appelés plans directeurs de la surface. 11 en existe deux, définis chacun à une translation près.

Les ellipsoïdes et les hyperboloïdes ont, à la fois, un centre et des plans de symétrie. Les paraboloïdes n'ont que des plans de symétrie.

4. Problèmes tangentiels

Les dix-sept variétés de quadriques donnent une idée assez complète des différentes formes que peuvent prendre les surfaces de l'espace usuel. Paraboloïde hyperbolique et hyperboloïde à une nappe fournissent des exemples très simples de surfaces qui traversent leurs plans tangents (ce qui n'est pas le cas de la sphère ou du cylindre de révolution, par exemple).

Une étude assez simple, fondée sur la théorie des valeurs et des vecteurs propres d'une matrice réelle symétrique, permet de déterminer les plans qui coupent une quadrique suivant des cercles. Ces plans sont parallèles entre eux (on se limite à des plans réels). Certains plans limites sont tangents à la surface en des points appelés *ombilics*. Il y en a deux pour un ellipsoïde réel non sphérique, deux pour un hyperboloïde à une nappe, deux pour un paraboloïde elliptique. Il existe des cas particuliers ; par exemple, **tous** les points d'une sphère sont des ombilics.

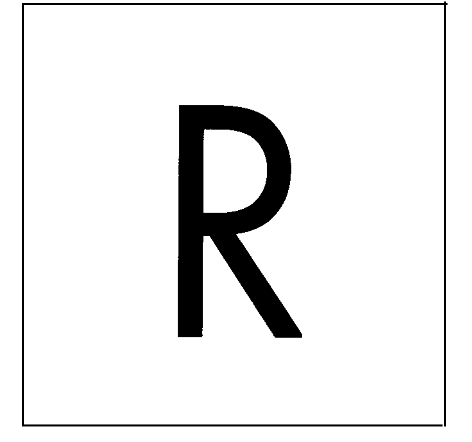
Les quadriques propres sont non seulement des ensembles de points soumis à des conditions du second degré, mais aussi des enveloppes de plans dont les paramètres annulent un polynôme homogène du second degré, autre forme quadratique attachée à la surface. Aussi dit-on que ces quadriques sont des enveloppes *de seconde classe*. Un cône, par exemple, ne répond pas à cette définition, car il possède deux équations tangentieiies au iieu d'une.

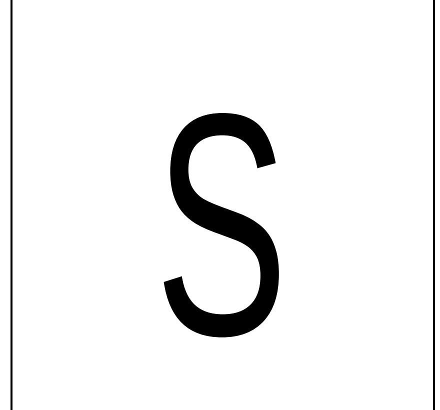
Les quadriques généralisent donc étroitement les propriétés affines et projectives des coniques (cf. CONIQUES). Il faut noter toutefois que, hormis les quadriques de révolution, obtenues par simple rotation d'une conique autour d'un axe, il n'existe pas de concept analogue à ceux de foyers et de directrices pour les coniques. Ces notions métriques sont donc directement liées à la structure particulière du plan. Sans doute cela provient-il, comme pour la plupart des résultats non généralisables si la dimension de l'espace excède deux, de la structure des rotations planes dont le groupe cesse d'être commutatif lorsque la dimension passe de deux à trois.

ANDRÉ WARUSFEL

Bibliographie

M. BERGER, *Géométrie* / t. IV, *Formes quadratiques, quadriques et coniques*, Cedic/Fernand Nathan, Paris, 1978 / G. CAGNAC, E. RAMIS & J. COMMEAU, *Traité de mathématiques spéciales*, t. *III : Géométrie*, Masson, Paris, 1967 / G. CAGNAC & H. COMMISSAIRE, *Cours de mathématiques supérieures et spéciales*, t. II, *ibid.*, 1951 / P. MARTIN, *Applications de l'algèbre* et de *l'analyse à la géométrie*, A. Colin, Paris, 1967 / A. WARUSFEL. *Dictionnaire raisonné de mathématiques* (pour la classification des dix-sept sortes de quadriques d'après les éléments de la matrice *A*), Seuil, Paris, 1966.





REPÈRE AFFINE → AFFINES ESPACE & REPÈRE

REPÈRE PROJECTIF → PROJECTIFS ESPACE & REPÈRE

SÉRIES & PRODUITS INFINIS

La notion de limite d'une suite est à la base de l'analyse. Le langage des séries, équivalent à celui des suites, s'est imposé dès le XVIIe siècle à propos du développement des fonctions en série entière. Cependant, les fondements rigoureux de la théorie des séries, reposant sur une définition des limites, remontent seulement au début du XIXe siècle, avec les travaux d'Abel, de Cauchy et de Gauss. L'étude des séries de nombres réels ou complexes et celle des séries de fonctions (séries entières, séries de Fourier, etc.) peuvent être considérées comme des cas particuliers de la théorie des séries d'éléments d'un espace vectoriel normé. On peut regrouper la notion de produit infini, utilisée par Euler au XVIIIe siècle, avec celle de série, à condition de se placer dans le cadre des groupes topologiques séparés.

Séries

Soit G un groupe commutatif topologique séparé (cf. algèbre TOPOLOGIQUE), dont la

loi est notée additivement. On appelle série d'éléments de G un couple A = $((u_n), (s_n))$ constitué de deux suites d'éléments de G telles que, pour tout entier naturel n, on ait :

$$(1) \qquad s_n = \sum_{m=0}^{n} u_m ;$$

l'élément s_n s'appelle *somme à l'ordre n*, la suite (u_n), *terme général*, et la suite (s_n), *suite des sommes partielles* de la série A.

On dit que la série A est *convergente ou divergente* suivant que la suite (s_n) converge ou non. Lorsque la série A est convergente, la limite s de (s_n) s'appelle *somme* de A et se note encore :

$$\sum_{n=0}^{+\infty} u_n ;$$

dans ces conditions, pour tout entier naturel n, l'élément $r_n = s - s_n$ s'appelle *reste à l'ordre n* et se note :

$$\sum_{m=n+1}^{+\infty} u_m.$$

Il est immédiat que, si la série A converge, son terme général tend vers 0. Examinons les liens entre suites et séries. Pour toute suite (u_n) d'éléments de G, il existe une série A et une seule dont le terme général est (u_n) ; sa somme à l'ordre n est définie par la relation (1). Inversement, pour toute suite (s_n) d'éléments de G, il existe une série A et une seule dont la suite des sommes partielles est (s_n) ; son terme général est défini par les relations :

$$u_0 = s_0,$$
$$u_n = s_n - s_{n-1}, \quad n \neq 0$$

Ainsi, par définition, l'étude de la convergence d'une série se ramène à celle d'une suite. Réciproquement, les règles de convergence des séries peuvent servir à étudier la convergence d'une suite par l'intermédiaire de la série des différences.

Le cas fondamental dans la théorie des séries est celui où G est le groupe sous-jacent à un espace vectoriel normé E. Les séries d'éléments de E constituent un espace vectoriel ; les séries convergentes constituent un sous-espace vectoriel de l'espace vectoriel précédent, et l'application qui à toute série convergente fait correspondre sa somme est linéaire. La multiplication de Cauchy des séries d'éléments d'une algèbre normée ne présente d'intérêt que dans le cas des séries entières ; nous n'indiquerons ici que la multiplication des familles sommables (cf. *infra).*

Lorsque l'espace vectoriel normé E est complet, le *critère de convergence de Cauchy* prend la forme suivante : Pour qu'une série A = $((u_n), (s_n))$ converge, il faut et il suffit que, pour tout voisinage V de 0, il existe un entier naturel n_0 tel que, pour tout couple (q, r) d'entiers naturels avec $r > q \geqslant n_0$, on ait :

$$\sum_{m=q+1}^{r} u_m \in \mathrm{V}.$$

tien avec les intégrales impropres

Supposons toujours l'espace vectoriel normé E complet. L'étude de la convergence d'une intégrale impropre peut se ramener à celle d'une série, et réciproquement. Soit en effet f une application réglée (cf. CALCUL INFINITÉSIMAL-Calcul à une variable, chap. 3) de $[0, +\infty[$ dans E admettant 0 pour limite à l'infini, (α_n) une suite strictement croissante de nombres réels positifs tendant vers $+\infty$ telle que $\alpha_0 = 0$ et que la suite $(\alpha_{n+1} - \alpha_n)$ soit

bornée, et enfin A la série dont le terme général est défini par la relation :

$$u_n = \int_{\alpha_n}^{\alpha_{n+1}} f(t)\, dt\,;$$

pour que l'intégrale impropre :

$$\int_0^{+\infty} f(t)dt$$

converge, il faut et il suffit que la série A converge.

Convergence des séries de nombres réels positifs

Dans le cas des séries de nombres réels positifs, on peut obtenir des règles plus précises de convergences des séries, grâce au résultat fondamental suivant : Pour qu'une série A $= ((u_n), (s_n))$ de nombres réels positifs converge, il faut et il suffit que la suite (s_n) soit majorée. Plus précisément, si cette suite est majorée, on a :

$$\sum_{n=0}^{+\infty} u_n = \sup_{n \in \mathbb{N}} s_n\,;$$

si cette suite n'est pas majorée, s_n tend vers $+\infty$.

Soit A et B deux séries de nombres réels positifs, de termes généraux (u_n) et (v_n). Si, pour tout entier naturel n, on a $u_n \leqslant v_n$, il découle du théorème ci-dessus que la convergence de la série B implique celle de la série A. Dans ce cas :

$$\sum_{n=0}^{+\infty} u_n \leqslant \sum_{n=0}^{+\infty} v_n.$$

Ce corollaire permet de ramener l'étude de la plupart des séries à celle de séries beaucoup plus simples, qui serviront alors de séries de référence pour les séries les plus générales. La convergence de ces séries de référence.s'établit en les comparant à des intégrales impropres, ce qui montre l'importance du résultat que voici.

Soit f une fonction réglée sur $[0, +\infty[$ à valeurs réelles positives, décroissante et ayant 0 pour limite à l'infini, et A la série de terme général $(f(n))$. Pour que la série A converge, il faut et il suffit que l'intégrale impropre :

$$\int_0^{+\infty} f(t)\, dt$$

converge.

On prend la plupart du temps pour séries de références les *séries géométriques,* c'est-à-dire les séries dont le terme général est de la forme (a^n), les *séries de Riemunn,* de terme général $(1/n^{\alpha})$, convergentes si et seulement si $\alpha > 1$, les *séries de Bertrand,* de terme général :

$$\frac{1}{n(\ln n)^{\alpha}},$$

convergentes également si et seulement si $\alpha > 1$.

La comparaison directe des séries de nombres réels positifs s'effectue à l'aide de la règle suivante : soit A et B deux séries de nombres réels positifs, de termes généraux (u_n) et (v_n). Si (u_n) est dominée par (v_n) et si, de plus, B converge, alors A converge. Il s'ensuit que, si (u_n) et (v_n) sont semblables, les séries A et B sont toutes deux convergentes ou toutes deux divergentes,

En prenant pour série de référence une série géométrique ou bien une série de Riemann, on obtient les règles classiques que voici.

Règle de Cauchy. Soit (u_n) une suite de nombres réels positifs telle que $\sqrt[n]{u_n}$ admette une limite β. Si $\beta < 1$, la série de terme général (u_n) converge ; si $\beta > 1$, cette série diverge.

Règle de Riemunn. Soit (u_n) une suite de nombres réels positifs telle qu'il existe un nombre réel α satisfaisant à la condition suivante : La suite $n^{\alpha}u_n$ admet une limite β

appartenant a $[0, +\infty[$. Si $\alpha > 1$ et si $\beta < +\infty$, la série de terme général (u_n) converge ; si $\alpha \leqslant 1$ et si $\beta \neq 0$, cette série diverge.

La comparaison directe s'utilise souvent sous la variante suivante, appelée comparaison logarithmique : soit (u_n) et (v_n) deux suites de nombres réels strictement positifs telles qu'a partir d'un certain rang on ait :

$$\frac{u_{n+1}}{u_n} \leqslant \frac{v_{n+1}}{v_n};$$

si la série de terme général (v_n) converge, il en est de même de la série de terme général (u,,).

En prenant toujours pour série de référence une série géométrique ou une série de Riemann, on obtient les deux autres règles classiques suivantes.

Règle de D'Alembert. Soit (u,,) une suite de nombres réels strictement positifs telle que u_{n+1}/u_n admette une limite β. Si $\beta < 1$, la série de terme général (u_n) converge : si $\beta > 1$, cette série diverge.

Règle de Ruube-Duhamel. Soit (u,) une suite de nombres réels strictement positifs telle que u_{n+1}/u_n tende vers 1 par valeurs inférieures. On considère la suite (t,,) définie par la relation :

$$t_n = 1 - \frac{u_{n+1}}{u_n};$$

on suppose enfin que nt_n admet une limite β appartenant à $[0, +\infty[$. Si $\beta > 1$, la série de terme général (u_n) converge ; si $\beta < 1$, cette série diverge.

Convergence absolue et semi-convergence

L'étude d'une série d'éléments d'un espace de Banach peut souvent se ramener à celle d'une série de nombres réels positifs, grâce à la notion suivante : On dit qu'une série $A = ((u_n), (s_n))$ d'éléments d'un espace vectoriel normé E est *absolument convergente* si la série de terme général $(\| u_n \|)$ est convergente. Pour que E soit complet, il faut et il suffit que toute série absolument convergente d'éléments de E soit convergente. En particulier, toute série absolument convergente de nombres complexes est convergente.

Prenons par exemple pour E l'espace vectoriel des applications bornées sur un ensemble X à valeurs dans un espace de Banach F, et munissons E de la norme de la convergence uniforme. La convergence absolue au sens de cette norme est dite *normale*. Toute série normalement convergente d'éléments de E est uniformément convergente sur X et absolument convergente (au sens de la norme sur F) en tout point de X ; une telle série converge simplement sur X. On notera que toutes les réciproques sont fausses.

L'étude des séries non nécessairement absolument convergentes est souvent facilitée par la règle suivante.

Règle d'Abel. Soit (α_n) une suite décroissante de nombres réels positifs convergeant vers 0 et (a,) une suite d'éléments d'un espace de Banach E. S'il existe un nombre réel positif β tel que, pour tout couple (q, r) d'entiers naturels avec $q < r$, on ait :

$$\left\| \sum_{m=q+1}^{r} a_m \right\| \leqslant \beta,$$

alors la série de terme général $(\alpha_n a_n)$ est convergente. De plus, pour tout entier naturel n, le reste à l'ordre n est majoré en norme par $\beta\alpha_{n+1}$.

On retrouve ainsi la condition suffisante de convergence des séries alternées : soit (u_n) une suite de nombres réels non nuis telle que la suite $((-1)^n u_n)$ soit de signe constant. Si la suite (u,,) tend vers 0 et si la

suite $(|u_n|)$ est décroissante, alors la série de terme général (u_n) est convergente, son reste à l'ordre n est majoré en valeur absolue par u_{n+1} et a le signe de u_{n+1}.

Il existe donc des séries convergentes sans être absolument convergentes, telles que la série harmonique alternée, de terme général $(-1)^n/n$, pour $n \geqslant 1$; de telles séries sont dites *semi-convergentes*.

Familles sommables

La définition de la somme d'une série repose sur le fait que l'ensemble des indices est N, et donc un ensemble canoniquement ordonné. Dans de nombreux problèmes, l'ordre des termes ne joue aucun rôle. Le besoin se fait aussi sentir de définir la somme d'une famille indexée par un ensemble I (non nécessairement dénombrable *a priori)*, indépendamment du choix d'une relation d'ordre dans 1.

Soit de nouveau G un groupe commutatif topologique séparé. On dit qu'une famille a = (u,), $i \in$ I, d'éléments de G est *sommable* s'il existe un élément s de G satisfaisant à la condition suivante : Pour tout voisinage V de 0, il existe une partie finie J_0 de 1 telle que, pour toute partie finie J de I contenant J_0, on ait :

$$s - \sum_{i \in J} u_i \in V\,;$$

un tel élément s est unique. On l'appelle somme de la famille a et on le note :

$$\sum_{i \in I} u_i.$$

Si 0 admet une base dénombrable de voisinages, le support de toute famille sommable est dénombrable (ce qui ne signifie pas que l'on doive se ramener systématiquement au cas ou I = N).

Soit maintenant E un espace de Banach. Le critère de Cauchy devient : Pour qu'une famille (u_i), $i \in$ I, d'éléments de E soit sommable, il faut et il suffit que. pour tout voisinage V de 0, il existe une partie finie J_0 de I telle que, pour toute partie finie K de I ne rencontrant pas J,,. on ait :

$$\sum_{i \in K} u_i \in V.$$

La notion de famille sommable est commutative. De manière précise, pour toute famille sommable (u_i), $i \in$ 1, d'éléments de G et pour toute permutation σ de I, la famille $(u_{\sigma(i)}$, $i \in$ I, est sommable, et :

$$\sum_{i \in I} u_{\sigma(i)} = \sum_{i \in I} u_i$$

Examinons le cas où I = N. Soit (u_n) une suite d'éléments de G. On dit que la série de terme général (u_n) est *commutativement convergente* si, pour **toute** permutation σ de N, la série de terme général $(u_{\sigma(n)})$ est convergente. Si la suite (u,,) est sommable, la série de terme général (u,,) est commutativement convergente. Réciproquement, dans le cas des espaces de Banach, la convergence commutative implique la sommabilité ; de plus, pour toute permutation σ de N,

$$\sum_{n=0}^{+\infty} u_{\sigma(n)} = \sum_{n=0}^{+\infty} u_n = \sum_{n \in N} u_n.$$

Soit (u,), $i \in$ I, une famille sommablc d'éléments d'un espace de Banach E et (I_h), $h \in$ H, une partition de I. Alors, pour tout élément h de H, la famille (u_i), $i \in I_h$, est sommable, la famille (v_h), $h \in$ H, où :

$$v_h = \sum_{i \in I_h} u_i,$$

l'est encore, et :

$$\sum_{i \in I} u_i = \sum_{h \in H} \left(\sum_{i \in I_h} u_i \right);$$

cette formule est dite formule de sommation par paquets. Soit E,, E_2, et F trois espaces de Banach et S une application bilinéaire continue de E, X E_2 dans F ; soit (u_i), $i \in I$, une famille sommable d'éléments de E, et (v,), $j \in$ J, une famille sommable d'éléments de E_2. Alors, la famille $(S(u_i, v_j))$, $(i, j) \in I \times J$, est sommable, et on a :

$$S\left(\sum_{i \in I} u_i, \sum_{j \in J} v_j \right) = \sum_{\substack{i \in I \\ j \in J}} S(u_i, v_j);$$

en particulier, on peut définir le produit de deux familles sommables d'éléments d'une algèbre de Banach.

La définition des familles absolument sommables d'éléments d'un espace vectoriel normé est calquée sur le cas des séries. Toute famille absolument sommable est sommable. La réciproque est vraie lorsque l'espace vectoriel E est de dimension finie (mais elle ne l'est pas si l'on suppose seulement que E est complet). en particulier, toute série absolument convergente de nombres complexes est commutativement convergente.

Séries multiples

La théorie des familles sommables s'applique notamment aux séries multiples. Étant donné un espace de Banach E, un entier naturel non nul r et une partie infinie I de $\mathbf{Z}^r$, on appelle série r-uple d'éléments de E indexée par I tout couple A = $((u_t), (s_J))$ constitué d'une suite r-uple d'éléments de E et d'une famille (s_J) d'éléments de E, où J parcourt l'ensemble des parties finies de I, telles que :

$$s_J = \sum_{t \in J} u_t;$$

la suite (u_t), $t \in I$, s'appelle terme général de la série A.

On dit qu'une telle série est absolument convergente si la famille (u_t), $t \in I$, est absolument sommable. La somme :

$$\sum_{t \in I} u_t$$

s'appelle alors comme la série A.

Prenons, par exemple, $I = \mathbf{Z}^r - \{0\}$, $I_+ = \mathbf{N}^r - \{0\}$, α un nombre réel non nul et, pour tout élément $t = (n_1, n_2, \dots n_r)$ de 1, posons :

$$u_t = \frac{1}{(|n_1| + \dots + |n_r|)^\alpha};$$

les séries r-uples de termes généraux (u_t), $t \in I$, et (u_t), $t \in I_+$, convergent si et seulement si $\alpha > r$. Plus généralement, pour toute norme sur l'espace vectoriel $\mathbf{R}^r$, les séries r-uples de termes généraux :

$$\left(\frac{1}{\|t\|^\alpha} \right), t \in I, \quad \left(\frac{1}{\|t\|^\alpha} \right), t \in I_+,$$

convergent si et seulement si $\alpha > r$.

Soit maintenant A une série double de terme général $(u_{n,p})$, où (n, p) parcourt $\mathbf{N}^2$. Si cette série est convergente, alors, pour tout entier naturel n, la série de terme général $(u_{n,p})$, $p \in$ N, est convergente, et la série de terme général (v_n) avec :

$$v_n = \sum_{p=0}^{+\infty} u_{n,p}$$

est convergente. De plus :

$$\sum_{n,p} u_{n,p} = \sum_{n=0}^{+\infty} \left(\sum_{p=0}^{+\infty} u_{n,p} \right),$$

dite formule de sommation par lignes des séries doubles. On peut énoncer de même une formule de sommation par colonnes. Les réciproques sont vraies si A est une série de nombres réels positifs.

Produits infinis

Soit G un groupe commutatif topologique séparé. Lorsque la loi de G est **notée** multiplicativement, les séries et les familles sommables d'éléments de G prennent respectivement les noms de produits infinis et de familles multipliables.

Cependant, lorsque le groupe G est le groupe multiplicatif d'un corps commutatif topologique séparé K, une suite d'éléments de G ne converge au sens de G que si elle converge vers un élément non nul de K. Cette remarque conduit à modifier légèrement les définitions.

On appelle produit infini d'éléments de K un couple A = ((u,), (p_n)) constitué de deux suites d'éléments de K telles que, pour tout entier naturel n, on ait :

$$p_n = \prod_{m=0}^{n} u_m ;$$

l'élément p_n s'appelle produit à l'ordre n de A, et la suite (u_n) s'appelle terme général de A.

On dit que le produit infini A est convergent dans K si u_n est non nul à partir d'un certain rang n_0 et si le produit infini de terme général (u_n), $n \geqslant n_0$, est convergent dans G = K*. Il est immédiat que le terme général d'un produit infini convergeant dans K converge vers 1.

On dit de même qu'une famille (u_i), $i \in$ I, d'éléments de K est multipliable dans K si le support I_0 de cette famille, c'est-à-dire l'ensemble des indices i tels que $u_i \neq 0$, est le complémentaire d'une partie finie de I et si la famille (u_i), $i \in I_0$, est multipliable dans K*.

Les produits infinis ayant leurs principales applications dans la théorie des fonctions analytiques, nous nous plaçons désormais dans le cas du corps des nombres complexes.

Le *critère de Cauchy* devient ici : Pour que le produit infini A converge, il faut et il suffit que, pour tout voisinage V de 1, il existe un entier naturel n_0 tel que, pour tout couple (q, r) d'entiers naturels tel que $r > q \geqslant n_0$, on ait :

$$\prod_{m=q+1}^{r} u_m \in V.$$

L'étude des produits infinis de nombres complexes se ramène à celle des séries : Soit (u_n) une suite de nombres complexes non réels négatifs ; pour que le produit infini de terme général (u_n) soit convergent (resp. commutativement convergent), il faut et il suffit que la série de terme général $(\ln u_n)$ soit convergente (resp. commutativement convergente). Soit (v_n) une suite de nombres complexes ; pour que le produit infini de terme général (1 + v,,) soit commutativement convergent, il faut et il suffit que la série de terme général (v_n) soit absolument convergente.

Comme le terme général d'un produit infini convergent (u_n) tend vers 1, on pose u_n = 1 + v,,, où v,, → 0. Lorsque (u_n) est une suite de nombres réels *positif,* la convergence du produit infini de terme général u_n équivaut à celle de la *série* de **terme** général v_n, par passage au logarithme, car $\ln(1 + x) \sim x$ au voisinage de 0.

On est alors amené à définir la convergence absolue d'un produit infini de terme général u_n par la convergence absolue de la série de terme général v_n. Tout produit absolument convergent est convergent.

LUCIEN CHAMBADAL

Bibliographie

M. BALABANE, M. DUFLO, M. FRISH et al., *Maths en kit*, t. II : *Sommes : séries*, Vuibert, Paris, 1982 / N. BOURBAKI, *Topologie générale*, chap. III à IX, Masson, Paris, nouv. éd. 1982 / L. CHAMBADAL & J.-L. OVAERT, *Analyse* II. Gauthier-Villars, Paris, 1972 / Y. CHEVALLARD & R. ROLLAND, *Théorie des séries*, 2 vol.. Cedic-Nathan, Paris, 1979 / J. COMBES, *Suites et séries*, P.U.F., Paris, 1982 / J. DIEUDONNÉ, *Éléments d'analyse*, t. I et II, Gauthier-Villars, 1962-1982 / G. H. HARDY, *Divergent Series*, Chelsea Publ., New York, 2' éd. 1991 / E. RAMIS, C. DESCAMPS & J. ODOUX, *Cours de mathématiques spéciales*, t. IV : *Séries, équations différentielles...*, Masson, 1993 / L. SCHWARTZ, *Analyse. Topologie générale et analyse fonctionnelle*, Hermann, 1970 / E. WHITTAKER & Ci. N. WATSON, *A Course of Modern Analysis*, Cambridge Univ. Press, New York-Londres, 1969.

SÉRIES TRIGONOMÉTRIQUES

Les séries trigonométriques se sont introduites au XVIIIe et au début du XIXe siècle, en liaison avec certains problèmes de physique (mouvement des cordes vibrantes, propagation de la chaleur). Elles sont d'un usage courant en astronomie, en cristallographie, en optique. Mais c'est en mathématiques qu'elles ont joué le rôle le plus important.

La justification du formalisme introduit par Joseph Fourier a occupé une grande part de l'effort des analystes du XIXe et même du XXe siècle. Elle a conduit au concept moderne de fonction, à la théorie de l'intégration, aux notions les plus importantes concernant ia sommation des séries et enfin à une partie de l'analyse fonctionnelle moderne. Il se trouve même qu'un problème concernant les séries trigonométriques est à l'origine de la théorie des ensembles. Les séries trigonométriques constituent donc l'exemple type d'un objet mathématique introduit par les besoins de la physique et dont l'étude a conduit à l'élaboration de concepts et de théories mathématiques de grande portée.

Ce rôle, sans être aussi important qu'autrefois, n'est pas terminé, et l'article s'efforcera d'en donner une idée.



1. Notations

Les séries trigonométriques sont les séries de la forme :

$$(1)\qquad \sum_{n=0}^{\infty}(a_n \cos n\omega t + b_n \sin n\omega t),$$

$$(2)\qquad \sum_{n=0}^{\infty} r_n \cos(n\omega t + \varphi_n),$$

dans lesquelles t désigne une variable réelle, ω un nombre > 0 (c'est la fréquence fondamentale), les a_n et les b_n des coefficients réels ($b_0 = 0$), les r_n des nombres $\geqslant 0$ (les amplitudes) et les φ_n des nombres réels définis modulo 2π (les phases). Les séries (1) et (2) sont liées par les formules :

$$r_n \cos \varphi_n = a_n, \qquad r_n \sin \varphi_n = -b_n.$$

Les sommes partielles s'écrivent :

$$S_N(t) = \sum_{n=0}^{N}(a_n \cos n\omega t + b_n \sin n\omega t)$$

$$= \sum_{n=0}^{N} r_n \cos(n\omega t + \varphi_n).$$

Il est souvent commode de les écrire :

$$(3)\qquad S_N(t) = \sum_{n=-N}^{N} c_n e^{in\omega t},$$

en posant $c_n = a_n + ib_n$, $n \geqslant 0$, et $c_n = a_n - ib_n$, $n \leqslant 0$. Cela amène à considérer, au lieu de séries (1) ou (2), des séries :

$$(4) \qquad \sum_{n=-\infty}^{\infty} c_n e^{in\omega t}, \quad \sum_{n \in \mathbf{Z}} c_n e^{in\omega t}$$

à coefficients c_n complexes, dont on définit encore les « sommes partielles » par (3).

C'est sous la forme inspirée de (4) que s'écrivent le plus commodément les « séries trigonométriques généralisées » :

$$(5) \qquad \sum_{n=-\infty}^{\infty} c_n e^{i\lambda_n t},$$

où la suite λ_n est réelle (les λ_n s'appellent les fréquences) et les « séries trigonométriques multiples » :

$$(6) \qquad \sum_{n \in \mathbf{Z}^k} c_n e^{i\omega(n,t)},$$

où $t = (t_1, t_2, ..., t_k) \in \mathbf{R}^k$ et où (n, t) est le produit scalaire $n_1t_1 + n_2t_2 + ... + n_kt_k$.

Dans la théorie des séries trigonométriques, on choisit généralement $\omega = 1$ (pour la commodité de l'écriture) ou $\omega = 2\pi$ (parce qu'alors les termes des séries (1), (2), (4) sont invariants par le changement de t en $t + 1$, et qu'ainsi t peut être considéré comme une variable sur le tore $\mathbf{T} = \mathbf{R}/\mathbf{Z}$, c'est-à-dire un nombre réel défini modulo 1). C'est ce dernier parti que nous prendrons.

Si f est une fonction, à valeurs complexes, définie sur T, c'est-à-dire une fonction périodique et de période 1, on pourra tenter de la représenter par une série trigonométrique. À cette fin, on lui associe la série (4), définie par :

$$(7) \qquad c_n = \int_{\mathbf{T}} f(t)\, e^{-2\pi int}\, dt, \qquad n \in \mathbf{Z};$$

on peut interpréter l'intégrale sur T comme une intégrale prise sur un intervalle quelconque de longueur 1. Si la fonction f est réelle, on peut aussi lui associer la série (1) définie par :

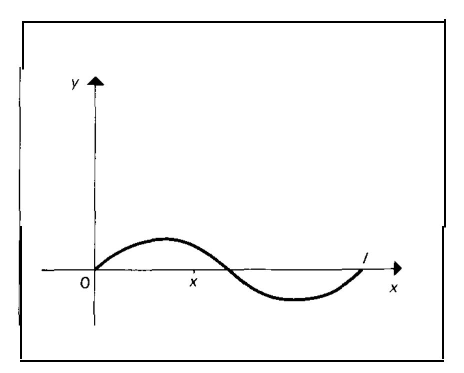
$$(8) \qquad \begin{cases} a_0 = \int_{\mathbf{T}} f(t)\, dt \\ a_n = \int_{\mathbf{T}} f(t) \cos 2\pi nt\, dt \\ b_n = \int_{\mathbf{T}} f(t) \sin 2\pi nt\, dt \end{cases}$$

où n est un entier $\geqslant 1$.

On appelle formules de Fourier les séries (7) et (8) ; leurs premiers membres s'appellent « coefficients de Fourier de f », et les séries (4), (1) ou (2) correspondantes « séries de Fourier de f ».

2. Aperçu historique

Quoique certaines sommes de séries trigonométriques aient déjà été calculées par L. Euler, on peut considérer que l'histoire des séries trigonométriques remonte à la solution, donnée par D. Bernoulli, du *problème des cordes vibrantes*. Le problème est de calculer (cf. figure) le mouvement d'une



corde, de longueur l, fixée en ses extrémités, et qui est soit écartée de sa position d'équilibre et lâchée (corde de guitare), soit frappée de façon à lui imprimer, en ses

différents points, des vitesses de déplacement latéral (corde de piano).

L'équation des cordes vibrantes, qui concerne le déplacement latéral $y(x, t)$ (supposé petit) au temps t du point x de la corde, est :

$$y''_{tt} = \omega^2 y''_{xx} ;$$

les conditions initiales imposent :

$$y(0, t) = y(l, t) = 0$$

et respectivement :

$$y(x, 0) = \varphi(x), \qquad y'_t(x, 0) = 0$$

pour la corde pincée, et :

$$y'_t(x, 0) = \psi(x)$$

pour la corde frappée. D'Alembert et Euler avaient découvert la solution générale, sous la forme :

$$y(x, t) = f(x + \omega t) - f(x - \omega t),$$

où f est une fonction périodique et de période $2l$ qui, dans le premier cas, est impaire et égale à $\varphi/2$ sur $[0, l]$ et, dans le second cas, est paire et primitive de $\psi/2\omega$ sur $[0, l]$. Pour des raisons physiques évidentes, D. Bernoulli pensait pouvoir écrire la solution sous la forme d'une série d'harmoniques solutions de l'équation des cordes vibrantes, c'est-à-dire :

$$y(x, t) = \sum b_n \sin \frac{nx}{l} \cos \frac{n\omega t}{l},$$

dans le premier cas et, dans le second cas,

$$y(x, t) = \sum \beta_n \sin \frac{nx}{l} \sin \frac{n\omega t}{l}.$$

Mais cela supposait, par exemple dans le premier cas, que l'on puisse écrire $\varphi(x)$ sous !a forme :

$$2 \sum b_n \sin \frac{nx}{l}$$

Comment une fonction $\varphi(x)$ arbitraire pourrait-elle se résoudre en une somme de fonctions sinus d'arcs multiples? Les meilleurs mathématiciens de l'époque (1750) ne le croyaient pas possible.

La question ne fut reprise que cinquante ans plus tard, par Fourier, à l'occasion de la *théorie analytique de la chaleur* (1822). L'équation en cause est ici :

$$y'_t = a y_{xx},$$

où y est la température au temps t et au point x d'une barre maintenue à température fixe (par exemple 0) aux extrémités, et une solution formelle en est :

$$y(x, t) = \sum b_n \exp\left(-an^2 \frac{t}{l^2}\right) \sin \frac{nx}{l},$$

de sorte que :

$$y(x, 0) = \sum b_n \sin \frac{nx}{l}.$$

De nouveau, on est amené à tenter d'écrire une fonction donnée sous forme d'une série trigonométrique. Fourier donne une série d'exemples, fondés sur des formules du type (8). 11 conclut, un peu rapidement, que les séries trigonométriques obtenues sont convergentes, qu'elles ont bien pour somme les fonctions données et qu'ainsi sont levées les objections faites à D. Bernoulli.

Il n'en est rien. Mais une bonne part de l'analyse mathématique allait sortir de cette intuition de Fourier.

L'étape décisive est l'admirable mémoire de 1829 où P. G. Lejeune-Dirichlet donne le premier *théorème de convergence de séries de Fourier*. Après avoir établi, pour une fonction f monotone et continue entre 0 et h, la formule :

$$\lim_{\lambda \to \infty} \int_0^b f(t) \frac{\sin \lambda t}{\sin t} dt = \frac{\pi}{2} f(0), \tag{9}$$

Dirichlet montre que, pour toute fonction fmonotone et continue par morceaux sur le tore T, les sommes partielles $S_N(t)$ de la série de Fourier de f convergent, en tout point t, vers :

$$\frac{1}{2}(f(t+0)+f(t-0)),$$

moyenne des valeurs limites de f à droite et à gauche de t. Cela résulte de (9) et de l'importante formule :

$$(10)\quad S_N(t) = \int_T f(t-s)D_N(s)\,ds = \int_T f(s)D_N(t-s)\,ds,$$

où :

$$D_N(s) = \sum_{-N}^{N} e^{2\pi ins} = \frac{\sin(2N+1)\pi s}{\sin \pi s}$$

Dans la dernière intégrale (10), D_N joue le rôle d'un noyau de convolution. On l'appelle, naturellement, le noyau de Dirichlet.

L'intérêt du travail de Dirichlet n'est pas seulement dans le résultat, ni dans la méthode qui est fort belle. On peut considérer que le concept moderne de *fonction* remonte à ce mémoire. Auparavant, une fonction était donnée soit par une expression analytique, soit par une représentation graphique. Au contraire, pour Dirichlet, la fonction n'est qu'une loi qui à chaque valeur x de la variable fait correspondre $f(x)$. Pour expliquer, par exemple, que les intégrales (8) n'ont de sens que pour certaines fonctions, Dirichlet considère une fonction φ égale à a pour x rationnel et à b pour x irrationnel, a étant différent de b. Avec le concept d'intégrale qu'on avait à l'époque, et qui allait être formalisé par Riemann, il s'agit en effet d'une fonction non intégrable.

La thèse de Riemann « Sur la possibilité de représenter une fonction par une série trigonométrique » a pour résultat principal un *théorème de localisation* qui s'exprime grossièrement ainsi : Si deux fonctions sont égales au voisinage d'un point, leurs séries de Fourier ont les mêmes propriétés en ce point. Elle introduit une méthode puissante. Cette méthode consiste à associer à une série trigonométrique (4), à coefficients tendant vers 0, la série deux fois formellement intégrée :

$$\sum_{n\neq 0} \frac{1}{(in\,\omega)^2} c_n e^{in\omega t}$$

(on suppose pour simplifier $c_0 = 0$), qui converge vers une fonction continue $F(t)$, et elle consiste ensuite à étudier la différence seconde :

$$(11)\quad \frac{F(t+h)+F(t-h)-2F(t)}{h^2} = \sum_{n\neq 0} c_n \left(\frac{\sin n\,\omega h}{n\,\omega h}\right)^2 e^{in\omega t},$$

quand $h \to 0$. C'est ce qu'on appelle le procédé de sommation de Riemann.

C'est dans la note historique qui précède la thèse que B. Riemann, critiquant une erreur commise par A. Cauchy, précise la théorie des séries numériques en distinguant les séries absolument convergentes (qui sont aussi commutativement convergentes) et les séries semi-convergentes (auxquelles on peut donner n'importe quelle somme par changement de l'ordre des termes). Et c'est au tout début de l'étude proprement dite que se trouve exposée la théorie de l'*intégrale de* Riemann, c'est-à-dire le premier concept d'intégrale mathématiquement élaboré. À ce stade enfin, pour la première fois, les formules de Fourier (7) ou (8) ont un sens parfaitement clair !

Il était naturel que le concept de fonction et celui d'intégrale apparussent à l'occasion de l'étude des séries de Fourier. La *théorie des ensembles* aurait pu naître autrement. Il se trouve qu'elle aussi a été fondée, par G. Cantor, pour poser et résoudre un problème sur les séries trigonométriques. Il s'agit maintenant de séries (1) qui ne sont pas nécessairement séries de Fourier. Si deux telles séries convergent en tout point vers la même fonction, sont-elles nécessairement identiques ? En d'autres termes, si la série (1) converge vers 0 pour tout t, a-t-on nécessairement $a_n = b_n = 0$ pour tout n ? Cantor répond affirmativement à la question. Puis, en 1871, sous le titre *Extension d'un théorème sur les séries trigonométriques,* il montre que le résultat subsiste si l'on suppose seulement que (1) converge vers 0 en dehors d'un ensemble fini ou, plus généralement, d'un ensemble dont un dérivé d'ordre fini ou transfini est vide. C'est loin d'être, dans cette direction, le meilleur résultat possible. Mais on peut voir là l'acte de naissance de la théorie des ensembles.

Deux illustres contre-exemples (1872-1873). K. Weierstrass donne, sous la forme d'une série :

$$(12) \qquad f(t) = \sum_0 a^n \cos b^n t,$$

où b est un entier $\geqslant 2$, où $0 < a < 1$ et où $ab \geqslant 10$, le premier exemple d'une fonction continue qui n'admet de dérivée en aucun point. Paul Du Bois-Reymond construit une fonction f continue, monotone par morceaux hors de tout intervalle de centre 0, mais oscillant indéfiniment au voisinage de 0, et dont la série de Fourier diverge au point 0.

L'année 1900 marque un tournant dans l'histoire des séries trigonométriques, avec les premiers travaux de L. Fejér et surtout la thèse de H. Lebesgue.

Fejér introduit les moyennes arithmétiques des sommes partielles $S_N(t)$, c'est-à-dire les :

$$\sigma_N(t) = \frac{1}{N}(S_0(t) + \dots + S_{N-1}(t)).$$

Ces « sommes de Fejér » s'expriment par une formule, analogue à (10),

$$(13) \quad \sigma_N(t) = \int_T f(t-s) K_N(s)\, ds = \int_T f(s) K_N(t-s)\, ds,$$

où :

$$K_N(t) = \frac{1}{N}(D_0(t) + \dots + D_{N-1}(t)) = \frac{\sin^2 \pi N t}{N \sin^2 \pi t}.$$

Fejér montre que $\sigma_N(t)$ tend vers :

$$\frac{1}{2}(f(t+0) + f(t-0))$$

chaque fois que cette expression a un sens et qu'en particulier $\sigma_N(t)$ tend vers $f(t)$ quand f est continue en t, uniformément si f est continue partout sur T. L'importance de ce résultat, en dehors de sa simplicité, est d'attirer l'attention sur la notion de *procédé de sommation.* À partir de là, il apparaît que, même si une série est divergente, il est raisonnable de lui attribuer une somme au moyen d'un procédé de sommation convenable.

L'idée n'était pas absolument nouvelle. Le procédé de Riemann, déjà décrit, consiste à associer à une série numérique :

$$\sum_0^\infty u_n,$$

l'expression, si elle existe :

$$\lim_{h \to 0} \sum_{0}^{\infty} u_n \left(\frac{\sin nh}{nh}\right)^2 ;$$

le procédé d'Abel-Poisson, qui s'introduisait dans l'étude des séries de Taylor, associe l'expression :

$$\lim_{r \uparrow 1} \sum_{0}^{\infty} u_n r^n ;$$

le procédé de Weierstrass, très lié à l'équation de la chaleur,

$$\lim_{\varepsilon \downarrow 0} \sum_{0}^{\infty} u_n e^{-\varepsilon n^2} ;$$

le procédé d'Émile Borel, introduit pour le prolongement analytique des séries de Taylor,

$$\lim_{\lambda \to \infty} \sum_{0}^{\infty} (u_0 + \dots + u_n) \frac{\lambda^n}{n!} e^{-\lambda}$$

et enfin le procédé de Fejér, ou procédé de Cesaro (C, 1),

$$\lim_{N \to \infty} \frac{1}{N} \sum_{0}^{N-1} (u_0 + \dots + u_n) = \sum_{0}^{N} \left(1 - \frac{n}{N}\right) u_n.$$

À partir de là s'imposait une étude systématique des procédés de sommation (0. Toeplitz, G. H. Hardy, M. Riesz...). L'application aux séries trigonométriques a donné lieu à une très vaste littérature, dont il faut particulièrement retenir ce qui concerne les séries trigonométriques multiples (6), dues à Salomon Bochner.

Les propriétés de la fonction K_N qui figure dans (13), et que l'on appelle le noyau de Fejér, expliquent le succès des σ_N ; ces propriétés sont partagées par d'autres fonctions dépendant d'un paramètre, par exemple le noyau de Poisson P, et le noyau de Weierstrass W_ε, qui permettent respectivement d'écrire :

$$\sum_{-\infty}^{\infty} c_n r^{|n|} e^{2\pi i n t} = \int_T f(t-s) P_r(s)\, ds,$$

$$\sum_{-\infty}^{\infty} c_n e^{-n^2 \varepsilon} e^{2\pi i n t} = \int_T f(t-s) W_\varepsilon(s)\, ds,$$

les c_n étant les coefficients de Fourier de f. Les propriétés en question, communes à K_N, P_r et W_ε, définissent les « identités approximatives » ; la convolution par une identité approximative est une bonne façon d'approcher une fonction. L'explicitation la meilleure de cette formule un peu vague se trouve dans le premier chapitre du livre de Y. Katznelson cité en référence (cf. représentation et approximation des FONCTIONS, chap. 4 et 7).

La thèse de Lebesgue (1902) donne une nouvelle définition de l'intégrale, plus générale que celle de Riemann, permettant par conséquent de donner un sens aux formules de Fourier (7) et (8) et à toutes celles qui en dérivent, pour une classe de fonctions beaucoup plus étendue. Plus tard, A. Denjoy, avec la totalisation, allait trouver une nouvelle généralisation de l'intégration, permettant d'attribuer un sens aux formules de Fourier pour toute fonction f susceptible de s'écrire comme somme, en tout point, d'une série trigonométrique partout convergente. *À chaque sens donné au symbole d'intégration correspond une définition des séries de Fourier* : on doit ainsi distinguer les séries de Fourier-Riemann, celles de Fourier-Lebesgue, celles de Fourier-Denjoy, celles de Fourier-Stieltjes (où dt est remplacé par une mesure sur T) et celles de Fourier-Schwartz (où $f(t)\, dt$ est remplacé par une distribution sur T). Dès 1906, dans ses *Leçons sur les séries trigonométriques,*

Lebesgue montrait le parti qu'on pouvait tirer de son intégrale, dans l'étude de la convergence ponctuelle et surtout dans l'étude de la convergence presque partout, qui renouvelait complètement le sujet. De 1906 date également la thèse de P. Fatou : *Séries trigonométriques et séries de Taylor,* où la notion de mesure de Lebesgue est largement utilisée.

L'importance en analyse de l'intégrale de Lebesgue s'est ainsi affirmée d'abord à l'occasion de la théorie des séries trigonométriques. Un exemple remarquable le fera comprendre. On connaissait depuis longtemps la « formule de Parseval » :

$$\int_{\mathrm{T}} |f(t)|^2 \, dt = \sum_{n \in \mathbf{Z}} |c_n|^2,$$

où les c_n sont les coefficients de Fourier de f. Cette formule vaut non seulement pour les fonctions intégrables au sens de Riemann, mais pour toutes les fonctions f de carré intégrable au sens de Lebesgue (ce que l'on écrit $f \in \mathrm{L}^2$). Un théorème, établi indépendamment par E. Fischer et par F. Riesz (1907), montre que toute suite $\{c_n\}$ telle que :

$$\sum_{n \in \mathbf{Z}} |c_n|^2 < \infty$$

(on écrit $\{c_n\} \in l^2$) est la suite des coefficients de Fourier-Lebesgue d'une certaine fonction de L^2 ; comme l'écrit quelque part F. Riesz, les formules de Fourier (7) sont comme un billet aller et retour entre L^2 et l^2 ; moins élégamment, on dit que c'est un isomorphisme d'espaces de Hilbert. Or, que L^2 soit un espace de Hilbert est un fait d'intérêt indépendant et de grande portée. Plus généralement, il apparaît que les espaces L^p, de fonctions de pième puissance intégrable au sens de Lebesgue, sont des espaces normés et *complets* pour $p \geqslant 1$; c'est ce fait fondamental que l'on désigne quelquefois aujourd'hui sous le nom de théorème de Fischer-Riesz.

Il est difficile de dresser, même en grandes lignes, l'histoire des séries trigonométriques au cours du XX[e] siècle. Nous retiendrons seulement quelques sujets. Mais il est bon d'indiquer que l'épanouissement des écoles polonaise et russe, dans les années 1920-1940, dans le domaine de l'analyse fonctionnelle et des probabilités est intimement lié à la fois à la théorie de l'intégrale de Lebesgue (très mal connue en France à l'époque) et aux problèmes soulevés par les séries trigonométriques (également ignorés en France, à l'exception de A. Denjoy, S. Mandelbrojt et R. Salem). Le traité de A. Zygmund (première édition en 1935, deuxième édition rééditée en 1988) est la référence essentielle, et il a joué un rôle de premier plan dans la formation de plusieurs générations d'analystes.

3. Quelques problèmes et autres développements

La convergence des séries de Fourier

Dirichlet avait établi que la convergence des séries de Fourier avait lieu pour des fonctions monotones et continues par morceaux, Du Bois-Reymond qu'elle n'avait pas nécessairement lieu pour des fonctions continues. Le théorème de Fischer-Riesz établit, quant à lui, que les sommes partielles de la série de Fourier d'une fonction $f \in \mathrm{L}^2$ tendent vers f dans l'espace L^2.

Jusqu'en 1966, on n'a pas su si la série de Fourier d'une fonction continue sur T converge nécessairement sur un ensemble non vide. À cette date, L. Carleson a montré que, pour toute $f \in \mathrm{L}^2$, la série de

Fourier de f converge vers $f(t)$ presque partout. La réponse à la question posée est donc positive, C'est le meilleur résultat possible, dans ce sens que, étant donné un ensemble de mesure nulle sur T, il existe une fonction continue dont la série de Fourier diverge sur cet ensemble.

Le théorème de Carleson vaut en remplaçant L^2 par L^p, avec $p > 1$ (R. Hunt, 1967). Il ne vaut pas pour L^1, puisque, dès 1926, on connaissait des fonctions de L^1 dont la série de Fourier diverge partout (A. N. Kolmogorov).

Essentiellement, pour les séries de Fourier à une variable, le problème de la convergence se trouve résolu avec les travaux de Carleson et Hunt.

Pour les séries de Fourier à plusieurs variables du type (6), avec $k \geqslant 2$, il faut définir ce qu'on appelle les sommes partielles avant de poser le problème de la convergence. Si l'on prend les sommes partielles « cubiques », définies comme la somme des termes pour lesquels :

$$\sup(|n_1|, |n_2|, ..., |n_k|) < R,$$

on a le résultat analogue au théorème de Carleson et Hunt (Charles Fefferman, Per Sjölin). Si l'on prend les sommes partielles « sphériques » définies par :

$$n_1^2 + n_2^2 + + n_k^2 < R^2,$$

il existe pour tout $p < 2$ une fonction de $L^p(T^k)$ dont la série de Fourier diverge presque partout (Fefferman, 1972) ; la situation est donc très différente pour $k = 1$ et pour $k \geqslant 2$; le problème reste ouvert pour $p \geqslant 2$, $k \geqslant 2$.

La convergence des séries trigonométriques

La convergence des séries trigonométriques est une question toute différente de la précédente : on considère ici des séries dont les coefficients ne sont pas nécessairement donnés par les formules de Fourier. Le problème est le suivant : Une fonction f étant donnée, existe-t-il une série trigonométrique qui converge presque partout vers f? On peut supposer f à valeurs finies ou infinies, mais on doit la supposer mesurable au sens de Lebesgue.

C'est un sujet étudié, depuis 1916, par D. Menchov et l'école russe. Menchov a d'abord étudié le cas $f \equiv 0$. Dans ce cas, le problème a évidemment une solution (la série à coefficients tous nuls), mais Menchov montre qu'elle n'est pas unique. Dans le cas où f a des valeurs finies, le problème a une solution positive. Le cas général est encore mystérieux. En particulier, on ne sait pas s'il existe une série trigonométrique dont les sommes partielles tendent vers $+\infty$ presque partout ; la réponse est vraisemblablement négative.

Si, au lieu de la convergence, on étudie la sommabilité d'Abel-Poisson, on obtient des résultats plus complets (N. Lusin et I. Privalov en 1925, F. Bagemihl et W. Seidel en 1954, J.-P. Kahane et aussi Y. Katznelson en 1971) : Étant donné deux fonctions f et g mesurables, à valeurs finies ou infinies, il existe une série trigonométrique (1) à coefficients tendant vers 0, qui est sommable vers f et dont la conjuguée :

$$\sum_{n=1}^{\infty} (a_n \sin n\omega t - b_n \cos n\omega t),$$

avec ici $\omega = 2\pi$, est sommable vers g.

Les ensembles d'unicité

Le problème remonte à Cantor. Quels sont les ensembles E sur la droite tels que, si une série trigonométrique (1) converge vers 0 en tout point n'appartenant pas à E, tous ses coefficients soient nuls ? On les appellera ensembles d'unicité. Tous les autres ensem-

bles sont appelés ensembles de multiplicité. Complétant les résultats de Cantor, W. H. Young montra que tout ensemble dénombrable est ensemble d'unicité. Dans l'autre sens, il est facile de voir que tout ensemble de mesure positive est ensemble de multiplicité. Le résultat de Menchov de 1916, qui a été indiqué plus haut, signifie qu'il y a des ensembles de mesure nulle qui sont ensembles de multiplicité.

La réunion de deux ensembles d'unicité fermés est un ensemble d'unicité (Nina Bari, 1927). Le résultat est faux pour des ensembles quelconques, inconnu pour des ensembles boréliens.

La classification des ensembles parfaits (fermés sans points isolés) en ensembles d'unicité et ensembles de multiplicité fait apparaître des phénomènes curieux (I. 1. Piatetski-Shapiro, R. Salem et A. Zygmund, A. Rajchman, R. Salem). L'ensemble triadique de Cantor est un ensemble d'unicité. Mais, si l'on considère un ensemble E_ξ parfait, décomposable en deux portions égales qui lui sont homothétiques dans le rapport ξ, avec $0 < \xi < 1/2$ (le cas $\xi = 1/3$ correspond à l'ensemble triadique de Cantor), la réponse dépend de propriétés arithmétiques du nombre $\theta = 1/\xi$; si θ est un entier algébrique dont tous les conjugués, à l'exception de lui-même, sont de module < 1 (on dit alors que θ est un nombre de Pisot), E_ξ est un ensemble d'unicité ; sinon, c'est un ensemble de multiplicité. Ce résultat est à la source de plusieurs travaux sur les nombres de Pisot et leurs généralisations (F. Bertrandias, Y. Meyer, J.-P. Schreiber).

Si l'on restreint l'ensemble des séries trigonométriques considérées, on agrandit la classe des ensembles d'unicité. Ainsi, si l'on donne une suite $\varepsilon_n \downarrow 0$, les ensembles E, tels que la seule série trigonométrique (1) vérifiant $|a_n| + |b_n| < \varepsilon_n$ et convergeant vers 0 hors de E soit la série à coefficients tous nuls, forment une classe $U(\varepsilon)$. Zygmund a montré qu'il existe des ensembles $U(\varepsilon)$ de mesure positive ; un problème ouvert est d'en trouver dont le complémentaire soit de mesure nulle.

Les séries trigonométriques absolument convergentes

Considérons les fonctions sommes de séries (4) absolument convergentes, avec $\omega = 2\pi$. On vérifie qu'en ajoutant, en soustrayant, en multipliant des fonctions de cette classe on reste dans la classe ; c'est dire que la classe en question est un anneau, qu'on désigne par A. Lorsqu'on munit chaque fonction de la norme $\sum |c_n|$, c'est un anneau normé. La théorie des anneaux normés est une des perles de l'analyse fonctionnelle ; elle est due à 1. M. Gelfand (1942). Mais, avant Gelfand, N. Wiener avait, en étudiant l'anneau A, dégagé certaines idées maîtresses de la future théorie. Le résultat principal (théorème de Wiener-Lévy) est que toute fonction analytique d'une fonction de la classe A est une fonction de la classe A ; en bref, les fonctions analytiques opèrent dans A.

La réciproque fut établie par Y. Katznelson en 1958 : Si F est une fonction de variable réelle qui opère dans A, cette fonction F est analytique. En 1959, P. Malliavin démontrait que les idéaux fermés dans A ne sont pas nécessairement déterminés par leur cospectre ; c'est-à-dire qu'une fonction $f \in A$ n'est pas nécessairement approchable dans A par des fonctions qui s'annulent au voisinage de l'ensemble de ses zéros.

Auparavant (en 1954), A. Beurling et H. Helson avaient montré que les seules applications φ de T dans T telles que $f \in A \Rightarrow f \circ \varphi \in A$ sont les fonctions linéaires $\varphi(t) = nt + a$.

Après ces résultats, l'intérêt principal s'est porté sur les classes A(E) de fonctions définies sur un ensemble fermé E C T et prolongeables en fonctions de la classe A. Ce sont de nouveaux anneaux normés, pour lesquels l'extension des théorèmes de Katznelson et de Malliavin n'est pas facile ; le meilleur résultat dans cette direction a été obtenu en 1965 par N. Varopoulos, au moyen de sa théorie des algèbres tensorielles : Les deux théorèmes s'étendent dès lors que E contient un ensemble de la forme E + F, les ensembles E et F étant deux fermés non dénombrables.

L'extension du théorème de Beurling et Helson est encore plus difficile ; elle a fait l'objet de travaux de N. Leblanc.

Pour certains ensembles E, la classe A(E) coïncide avec l'ensemble des fonctions continues sur E : on les appelle ensembles de Helson. Deux problèmes posés depuis plus de quinze ans ont été résolus tout récemment à leur propos. La réunion de deux ensembles de Helson est un ensemble de Helson (S. Drury et N. Varopoulos, 1971). Il existe un ensemble de Helson qui porte une distribution, non nulle, dont les coefficients de Fourier tendent vers 0 à l'infini (T. Körner, 1972) ; un théorème de Helson affirme qu'une telle distribution ne peut être une mesure. Le problème principal qui reste est la « conjecture de dichotomie » : Ou bien E est un ensemble de Helson, et toutes les fonctions continues opèrent dans A(E), ou bien seules les fonctions analytiques opèrent dans A(E).

Les séries trigonométriques lacunaires

Les séries trigonométriques lacunaires apparaissent pour la première fois dans l'exemple de Weierstrass, d'une fonction continue nulle part dérivable. Une fonction telle que (12) est très irrégulière ; mais elle manifeste une sorte de régularité dans l'irrégularité, bien mise en évidence par G. Freud (1962) et plus encore par M. Bruneau (1970).

Considérons une série (5), où $\{\lambda_n\}$ est une suite symétrique, c'est-à-dire telle que A $_n$ = — A,, satisfaisant à la condition de lacunarité : $A_{n+1} \geqslant q\lambda_n$, avec 4 > 1 et n = 1, 2, . .., qu'on appelle condition de Hadamard. Dans ce cas, les fonctions :

$$c_n e^{i\lambda_n t} + c_{-n} e^{-i\lambda_n t}$$

sont presque indépendantes, et on peut obtenir pour ces séries l'analogue de beaucoup de résultats concernant les sommes de variables aléatoires indépendantes ; il est d'ailleurs intéressant de noter que, dans certains cas, les résultats sur les séries trigonométriques lacunaires ont précédé ceux qui concernent les variables indépendantes, pourtant essentiellement plus simples.

Exemples

1. Si $\Sigma \ c_n|^2 < \infty$, la série (5) converge presque partout (Kolmogorov, 1924) ; inversement, si la série converge sur un ensemble de mesure positive, on a $\Sigma|c_n|^2 < \infty$, et de plus la somme $f(t)$ vérifie exp $(\lambda f^2) \in L^1$ pour tout $\lambda > 0$ (Zygmund).

2. Si la série (5) converge en tout point d'un intervalle, on a $\Sigma| \ c_n \ < \infty$ (S. Sidon).

Un problème général, posé par S. Mandelbrojt, est le suivant. La suite $\{\lambda_n\}$ étant donnée, supposons que f ait une série de Fourier de la forme (5) et quefsatisfasse à une propriété P sur un intervalle, arbitrairement petit. S'ensuit-il que f ait la même propriété partout ? Lorsque

$\lambda_{n+1} - \lambda_n \rightarrow \infty$, avec $n \rightarrow \pm \infty$, la réponse est positive pour de nombreuses propriétés P, par exemple :

a) la nullité,

b) l'appartenance à L^2,

c) l'appartenance locale à A,

d) la dérivabiiité d'ordre infini,

e) l'analyticité.

De plus, pour les propriétés (a), (h) et (c), on connaît explicitement les conditions nécessaires et suffisantes sur $\{\lambda_n\}$ pour que la réponse soit positive.

Si l'on prend pour P la propriété d'être bornée, ou continue, le problème devient plus difficile ; les progrès dans cette direction, très liée à la théorie des nombres, sont dus à Y. Meyer.

Les séries trigonométriques oléotoires

Les séries trigonométriques aléatoires sont des séries de la forme (1), où les a_n et les b_n représentent des variables aléatoires ; le cas le plus **simple est :**

$$a_n = a_n(\omega) = \pm \alpha_n, \qquad b_n = b_n(\omega) = + \beta_n$$

(variables aléatoires de Rademacher indépendantes) ; un autre cas important est :

$$a_n(\omega) = X_n(\omega)\alpha_n, \qquad b_n(\omega) = Y_n(\omega)\beta_n,$$

où les X_n et Y_n, sont des variables aléatoires gaussiennes centrées, normalisées et indépendantes.

Les séries du premier type apparaissent pour la première fois dans des notes de R. Paley et A. Zygmund (1932). Un de leurs résultats marquants est le suivant. Si l'on a :

$$\sum_1^{\infty} \alpha_n^2 = \infty,$$

il est presque sûr que la série :

$$\sum_1 \pm \alpha_n \cos 2\pi nt$$

n'est pas une série de Fourier-Lebesgue. Ainsi, aucune condition sur les amplitudes meilleure que la condition de Fischer-Riesz ne garantit qu'une série trigonométrique est une série de Fourier.

Les séries du second type ont été introduites par N. Wiener dans l'étude du mouvement brownien ; on a dans ce cas :

$$\alpha_n = \beta_n = \frac{1}{n}$$

Pour les unes et les autres, la plupart des propriétés intéressantes de la série trigonométrique aléatoire ont pour probabilité 0 ou 1. La probabilité est la même pour que :

a) la série converge presque partout.

b) elle soit une série de Fourier-Lebesgue,

c) elle représente une fonction appartenant à tous les L^p, avec $1 \leqslant p < \infty$ (PaleyZygmund).

La probabilité est la même pour que :

a') la série converge partout,

h') elle représente une fonction bornée,

c') elle représente une fonction continue (P. Billard. 1961).

4. Applications des séries trigonométriques

En mathématique, les séries trigonométriques n'ont cessé, depuis deux cents ans, de suggérer de nouveaux concepts et de nouveaux sujets d'étude. Sans occuper. dans la mathématique du XXe siècle, la place qu'elles tenaient au XIXe siècle, on

peut penser que leur influence n'est pas terminée.

Les méthodes fondées sur les sommes trigonométriques jouent un rôle important en théorie des nombres : problèmes de Goldbach et de Waring, répartition modulo 1. Le lien entre séries trigonométriques et séries de Taylor explique leur intérêt dans l'étude du comportement des fonctions analytiques à la frontière. Les séries trigonométriques généralisées, qui interviennent dans la théorie des fonctions presque périodiques, ont aussi été appliquées à la fonction $\zeta(s)$ de Riemann. On pourrait poursuivre la liste des exemples.

Nées avec le problème des cordes vibrantes et la théorie analytique de la chaleur, les séries trigonométriques ont conservé avec la physique un lien permanent, en particulier en optique, en astronomie et en cristallographie.

La mise en œuvre de programmes de transformées de Fourier rapides permet le traitement sur ordinateurs de données autrefois inexploitables. En un mot, les formules de Fourier, dans lesquelles les intégrales sont remplacées par des sommes finies pour se prêter au calcul, permettent le calcul de N coefficients au moyen d'un nombre d'opérations (additions, multiplications) qui est de l'ordre de N^2. La « transformée de Fourier rapide » permet d'obtenir ces N coefficients au moyen de $N \ln N$ opérations. Le gain est considérable.

C'est ainsi qu'en 1970, dans certains programmes du Centre interrégional de calcul électronique (C.I.R.C.É.) à Orsay, on pouvait calculer plus d'un million de coefficients en moins de dix minutes. Ces programmes ont été utilisés particulièrement en astrophysique.

JEAN-PIERRE KAHANE

Bibliographie

N. K. BARI, *Trigonometričeskie rjady*, Moscou, 1961 (trad. angl. M. F. Mullins : *A Treatise on Trigonometrics Series*, 2 vol., Macmillan, New York, 1965) / R. E. EDWARDS, *Fourier Series. A Modern Introduction*, vol. II. Springer, New York, 2ᵉ éd. 1982 / J.-P. KAHANE, *Séries de Fourier absolument convergentes*, *ibid.*, 1970 / Y. KATZNELSON, *Introduction to Harmonic Analysis*, Dover, New York, 2ᵉ éd. 1976 / H. LEBESGUE, *Leçons sur les séries trigonométriques*, A. Blanchard, Paris, 1975 / R. SALEM, *Œuvres mathématiques*, Hermann, Paris, 1967 / E. M. STEIN & G. WEISS. *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton Univ. Press, Princeton (N. J.), 1971 / N. WIENER, *The Fourier Integral and Certain of Its Applications*, Cambridge Univ. Press, New York, 1989 / A. ZYGMUND, *Trigonometric Series*, 2 vol., Cambridge Univ. Press, Cambridge (G.-B.), 1959, 2ᵉ éd. 1988.

SPECTRALE THÉORIE

L'objet de la théorie spectrale est d'obtenir, pour certains endomorphismes d'un espace hilbertien, des formes réduites analogues aux formes canoniques de Jordan pour les endomorphismes d'un espace vectoriel de dimension finie et aux formes diagonales pour les endomorphismes hermitiens d'un espace vectoriel hermitien de dimension finie. La théorie des applications de Hilbert-Schmidt, rencontrées pour la première fois à propos des équations intégrales, permet de construire une première généralisation des résultats obtenus en dimension finie. En fait, le cadre naturel de cette généralisation est celui des applications compactes, étudiées par F. Riesz.

Néanmoins, le cas des endomorphismes les plus généraux échappe à ce cadre ; il fait l'objet de la théorie spectrale de Hilbert, qui utilise les techniques de l'intégration. On a axiomatisé la théorie spec-

trale, grâce aux concepts généraux de C*-algèbre et d'algèbre hilbertienne.



1. Théorie spectrale algébrique

Tant en algèbre qu'en analyse, on est fréquemment amené à définir et à calculer des fonctions d'un endomorphisme u d'un espace vectoriel E sur un corps commutatif K (inverse, puissances, exponentielle, etc.). À cet effet, il est utile de chercher les droites de E stables par u. On est ainsi conduit aux notions de valeur propre et de vecteur propre. On dit qu'un élément non nul x de E est un *vecteur propre* de u si la droite engendrée par x est stable par u, c'est-à-dire s'il existe un élément λ de K tel que $u(x) = \lambda x$. On dit qu'un scalaire λ est unc *valeur propre de* u si le noyau de $u \quad \lambda I_E$ est non réduit à $\{0\}$ L'ensemble des valeurs propres de u s'appelle *spectre ponctuel* de u et se note sp (u).

Même lorsque E est de dimension finie et que K est algébriquement clos, il peut arriver que E ne soit pas somme directe de droites stables par u. C'est le cas par exemple lorsque u est un endomorphisme nilpotent non nul de E. On voit apparaître l'intérêt de la notion *d'endomorphisme diagonalisable* : on appelle ainsi un endomorphisme u de E tel que E soit somme directe de droites stables par u, ou encore tel qu'il existe une base de E constituée de vecteurs propres de u. Lorsque E est de dimension finie, cela revient à dire qu'il existe une base de E telle que la matrice associée à u dans cette base soit diagonale. Il peut arriver que plusieurs droites stables correspondent à une même valeur propre.

C'est pour cela que, pour toute valeur propre A de u, on introduit le *sous-espace propre* associé à A, à savoir le noyau de $u - \lambda I_E$. La somme des sous-espaces propres de u est toujours directe ; pour que u soit diagonalisable, il faut et il suffit que cette somme soit égale à E.

Dans le cas où u n'est pas diagonalisable, il convient d'introduire des sous-espaces vectoriels de E stables par u « plus gros » que les sous-espaces propres : on appelle *sous-espace spectral* de u associé à une valeur propre A de u le sous-espace vectoriel F_λ réunion des sous-espaces vectoriels :

$$E_{\lambda,r} = \text{Ker } [(u - \lambda I_E)^r],$$

où $r \in \mathbb{N}$. Si la suite $(E_{\lambda,r})$ est stationnaire, on dit que A est l'indice fini. Le plus petit des entiers r tels que $E_{\lambda,r} = F_\lambda$ s'appelle alors indice de A et se note n(A). Lorsque F_λ est de dimension finie, on dit que λ est de multiplicité finie ; la dimension de F_λ s'appelle alors multiplicité de la valeur propre A. La somme des sous-espaces spectraux de u est toujours directe ; on dit que u est *trigonalisable* si cette somme est égale à E.

Lorsque E est de dimension finie, le spectre de u est fini ; il est constitué des scalaires A tels que :

$$\det(\lambda I_E - u) = 0,$$

c'est-à-dire des racines du polynôme det (XI, $- u$), appelé *polynôme caractéristique* de u. De plus, toute valeur propre λ de u est de multiplicité finie et égale à la multiplicité de ia racine A du polynôme caractéristique de u. En outre, l'idéal de K[X] constitué des polynômes P tels que P(U) = 0 est non réduit à $\{0\}$; il admet donc un générateur, appelé *polynôme minimal* de u. Toute valeur propre A de u est d'indice fini et égal à la multiplicité de

la racine A du polynôme minimal de u. Enfin, le polynôme minimal de u divise le polynôme caractéristique de u ; ce résultat s'appelle *théorème de Hamilton-Cayley*.

Grâce à ces notions, on peut caractériser les endomorphismes diagonalisables et les endomorphismes trigonalisables. Pour que u soit trigonalisable, il faut et il suffit que le polynôme minimal (ou le polynôme caractéristique) de u soit scindé, c'est-à-dire décomposable en produit de facteurs du premier degré. Pour que u soit diagonalisable, il faut et il suffit que le polynôme minimal de u soit scindé et que toutes ses racines soient simples. D'autre part, pour que u soit trigonalisable, il faut et il suffit qu'il existe une base de E telle que la matrice associée à u dans cette base soit trigonale supérieure.

En combinant les caractérisations données ci-dessus, on obtient un résultat plus précis : soit u un endomorphisme trigonalisable ; pour toute valeur propre λ de u, il existe une base B_λ du sous-espace spectral F_λ telle que la matrice associée dans cette base à l'endomorphisme u_λ de F_λ coïncidant avec u soit de la forme :

$$M_\lambda = \begin{pmatrix} \lambda & & & * \\ & \lambda & & \\ & & \ddots & \\ 0 & & & \lambda \end{pmatrix}$$

La matrice R associée à u dans la base B obtenue en réunissant les bases B_λ est une matrice diagonale de matrices trigonales :

$$R = \begin{pmatrix} M_{A_1} \; 0 \; 0 \; M_{\lambda_2} 0 \; . \; \cdot \; \cdot \; M_{\lambda_r} \end{pmatrix}$$

Une telle matrice trigonale supérieure est dite *réduite*.

On peut enfin mettre la matrice associée à un endomorphisme trigonalisable sous une forme canonique, grâce à la notion de *matrice de Jordan* ; on appelle ainsi une matrice carrée de la forme :

$$\begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ 0 & 0 & \lambda & \dots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & & \lambda \end{pmatrix}$$

c'est-à-dire une matrice dont les éléments diagonaux sont égaux, dont les éléments juste au-dessus de la diagonale sont égaux à 1 et dont les autres éléments sont nuls.

Pour toute valeur propre A de l'endomorphisme trigonalisable u, il existe une base B_λ de F_λ telle que la matrice M_λ associée à u,, dans cette base soit une matrice diagonale de matrices de Jordan. La matrice J associée à u dans la base B est encore une matrice diagonale de matrices de Jordan ; en particulier, ses éléments juste au-dessus de la diagonale sont égaux à 0 ou à 1. Une telle matrice est dite *forme réduite de Jordan.* Le calcul des puissances successives de J s'effectue aisément à partir de la formule du binôme de Newton.

Soit par exemple u l'endomorphisme de $\mathbf{R}^5$ canoniquement associé à la matrice :

$$M = \begin{pmatrix} 1 & 1 & -1 & 2 & -1 \\ 2 & 0 & 1 & -4 & -1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 \\ 0 & 0 & -3 & 3 & -1 \end{pmatrix}$$

Le polynôme caractéristique de u est $(X+1)^2(X-1)^3$. L'endomorphisme u est

trigonalisable, mais il n'est pas diagonalisable. Soit B la base de $\mathbf{R}^5$ définie à partir de la base canonique par la matrice de passage :

$$P = \begin{pmatrix} 1 & 1/3 & 1 & 0 & 3/2 \\ -1 & -1/3 & -1 & 1 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1/3 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

La matrice $J = P^{1} MP$ associée à u dans la base B est sous la forme réduite de Jordan :

$$J = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -10 & & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

2. Théorie de Riesz des applications linéaires compactes

Applications linéaires compactes

Historiquement, la notion d'application linéaire compacte s'est introduite sous le nom d'application complètement continue : étant donné deux espaces vectoriels normés E et F, une application linéaire u de E dans F est dite *complètement continue* si de toute suite bornée (x_n) d'éléments de E on peut extraire une suite (y,,) telle que la suite $(u(y_p))$ soit convergente dans F.

F. Riesz fut le premier à remarquer que cette condition permet de retrouver tous ies résuitats de ia *théorie de Fredholm* (cf. équations INTÉGRALES, chap. 5). En utilisant la caractérisation des espaces métriques compacts à l'aide de la condition de Bolzano-Weierstrass, on voit immédialement qu'une application linéaire u de E dans F est complètement continue si et seulement si l'image par u de la boule unité de E est une partie relativement compacte de F. Sous cette forme, la notion d'application complètement continue peut se généraliser aux espaces vectoriels topologiques.

Plus précisément, soit E et F deux espaces vectoriels topologiques localement convexes séparés. On dit qu'une application linéaire u de E dans F est *compacte* (resp. *précompacte*) s'il existe un voisinage V de 0 dans E tel que $u(V)$ soit une partie relativement compacte (resp. précompacte) de F.

Toute application compacte est précompacte ; la réciproque est vraie si l'espace vectoriel F est complet ou, plus généralement, si toute partie fermée bornée de F est complète. Toute application précompacte est continue ; la réciproque est fausse. Ainsi, pour que l'application identique 1, de E soit précompacte, il faut et il suffit que E soit de dimension finie, auquel cas elle est compacte (lemme de F. Riesz).

Les applications compactes de E dans F constituent un sous-espace vectoriel de l'espace vectoriel des applications linéaires continues de E dans F.

Soit u une application linéaire continue de E dans F et v une application linéaire continue de F dans un troisième espace G. Si l'une des applications u et v est compacte, il en est de même de l'application composée $v \circ u$. En particulier, les endomorphismes compacts de E constituent un idéal bilatère de l'algèbre $\mathcal{L}(E)$ des endomorphismes continus de E.

Soit E' et F' les duaux topologiques de E et de F, munis de la topologie de la convergence uniforme sur les disques compacts. Alors, si u est compact, il en est de même de ${}^t u$.

Enfin, toute application linéaire de rang fini est compacte.

On peut énoncer des propriétés analogues pour les applications précompactes.

Revenons au cas particulier où les espaces E et F sont normés ; munissons l'espace vectoriel C(E, F) de la norme des applications linéaires continues, à savoir la norme de la convergence uniforme sur la boule unité de E. Alors les applications précompactes de E dans F constituent un sous-espace vectoriel fermé de $\mathcal{L}(E, F)$. Il en est de même des applications compactes de E dans F, lorsque F est complet. Il en résulte que la limite en norme d'une suite d'applications de rang fini est une application compacte. Réciproquement, lorsque l'espace vectoriel F est hilbertien, toute application compacte de E dans F est limite d'une suite d'applications de rang fini. Lorsque F est un espace de Banach, cette réciproque se ramène au problème suivant : l'application identique d'un espace de Banach est-elle limite forte de projecteurs de rang fini (propriété d'approximation) ? Ces deux problèmes ont été résolus par la négative en 1976.

Supposons maintenant que E et F sont des espaces de Banach ; soit E' et F' les duaux topologiques de E et de F, munis des normes correspondantes. Pour qu'une application linéaire continue u de E dans F soit compacte, il faut et il suffit que sa transposée ${}^tu$ soit une application compacte de F' dans E'.

En particulier, si E et F sont des espaces hilbertiens, l'adjointe u^* d'une application compacte u de E dans F est une application compacte de F dans E.

Les propriétés des applications de rang fini se généralisent aux applications linéaires compactes, ce qui fait le principal intérêt de ces dernières. Laurent Schwartz a dégagé le résultat fondamental suivant :

Théorème de finitude. Soit E et F deux espaces vectoriels localement convexes séparés, u et v deux applications linéaires continues de E dans F. On suppose que u est un isomorphisme de E sur F et que v est compacte. Alors le noyau de $w = u + v$ est de dimension finie, l'image de w est un sous-espace vectoriel fermé de codimension finie dans F, et on a la formule :

$$\dim \operatorname{Ker}(w) = \operatorname{codim}_F \operatorname{Im}(w),$$

dite formule du rang. De plus, w est un morphisme strict, c'est-à-dire que w définit un isomorphisme de $E/\operatorname{Ker}(w)$ sur $\operatorname{Im}(w)$.

Ce théorème a trouvé des applications non seulement dans la théorie spectrale, que nous esquissons ci-dessous, mais aussi dans la théorie des faisceaux analytiques cohérents (cf. GÉOMÉTRIE ALGÉBRIQUE, chap. 6) et dans celle des indices topologiques.

Spectre d'un endomorphisme compact

Examinons maintenant le cas particulier où E = F et supposons que le corps de base est le corps C des nombres complexes. On appelle spectre d'un endomorphisme continu u de E l'ensemble, noté $\operatorname{sp}(u)$, des nombres complexes A tels que $u - $ AI, ne soit pas inversible dans l'algèbre unitaire $\mathcal{L}(E)$. Les éléments de $\operatorname{sp}(u)$ s'appellent valeurs spectrales de u. Lorsque E est un espace de Banach, le spectre de u est une partie compacte non vide de C. Toute valeur propre de u est une valeur spectrale de u, mais la réciproque est fausse si E n'est pas de dimension finie ; il peut même arriver que u n'ait aucune valeur propre. C'est la principale raison

pour laquelle la réduction des endomorphismes continus de E nécessite des outils radicalement nouveaux (cf. *infra*, chap. 3). Cependant, lorsque l'endomorphisme u est compact, les principaux résultats de la réduction des endomorphismes d'un espace vectoriel de dimension finie se généralisent.

Plus précisément, soit u un endomorphisme compact d'un espace vectoriel localement convexe séparé E et λ une valeur spectrale non nulle de u.

a) Le nombre complexe λ est une valeur propre de u.

b) Le sous-espace propre :

$$E_\lambda = \mathrm{Ker}(u - \lambda I_E)$$

est de dimension finie. Plus généralement, pour tout entier naturel non nul Y, le sous-espace :

$$E_{\lambda,r} = \mathrm{Ker}\ (u - AI,)'$$

est de dimension finie. En outre, λ est une valeur propre d'indice fini. En particulier, le sous-espace spectral F_λ est de dimension finie.

c) Le sous-espace vectoriel :

$$E', = \mathrm{Im}(u - \lambda I_E)$$

est fermé de codimension finie dans E. Plus généralement, pour tout entier naturel non nul r,

$$E'_{\lambda,r} = \mathrm{Im}\ (u - \lambda I_E)^r$$

est fermé de codimension finie. 11 en est de même de :

$$F'_\lambda = \bigcap_{r=0}^{+\infty} E'_{\lambda,r}.$$

d) Les sous-espaces vectoriels F_λ et F', sont supplémentaires topologiques dans E (c'est-à-dire qu'ils sont supplémentaires et que les projecteurs associés sont continus). De plus, l'endomorphisme de F_λ coïncidant avec u est nilpotent, tandis que l'endomorphisme de F', coïncidant avec u est un automorphisme de l'espace vectoriel localement convexe séparé F'_λ.

En outre, le spectre d'un endomorphisme compact u d'un espace vectoriel localement convexe séparé E est une partie compacte non vide de C, et tout point de sp(u) autre que 0 est isole. Autrement dit, ou bien le spectre de u est fini, ou bien il est constitué de 0 et d'une suite (A,) de nombres complexes non nuls convergeant vers 0. Lorsque l'espace vectoriel E n'est pas de dimension finie, 0 appartient toujours au spectre de u. On notera néanmoins que les résultats précédents ne s'appliquent pas à la valeur spectrale 0. Or, il peut arriver que u soit compact et injectif et que le spectre de u soit réduit à $\{0\}$, ce qui signifie que u est quasi nilpotent, c'est-a-dire que :

$$\lim_{n \to +\infty} \|u^n\|^{1/n} = 0;$$

c'est le cas pour l'endomorphisme de $\mathcal{C}([0, 1])$ qui à toute fonction continue sur $[a, b]$ associe sa primitive s'annulant au point a.

Enfin, I'endomorphisme transposé ${}^t u$ d'un endomorphisme compact u a le même spectre que u et les mêmes valeurs propres non nulles. avec les mêmes indices et les mêmes multiplicités pour u et ${}^t u$. Plus précisément, pour tout entier naturel non nul r, le sous-espace :

$$\mathrm{Im}[({}^t u \quad \lambda I_{E'})^r]$$

est l'orthogonal de $\mathrm{Ker}[(u - \lambda I_E)^r]$ et de même le sous-espace :

$$\mathrm{Im}[(u - \lambda I_E)^r]$$

est l'orthogonal de $\mathrm{Ker}[({}^t u \quad \lambda I_{E'})^r]$

Alternative de Fredholm

Soit E un espace vectoriel de dimension finie, E* son dual, w un endomorphisme de E et ${}^t w$ le transposé de w. Alors l'alternative suivante est vraie (cf. algèbre LINÉAIRE ET MULTILINÉAIRE) :

a) Ou bien w et ${}^t w$ sont des automorphismes. Dans ce cas, pour tout élément y de E et pour tout élément y* de E*, les deux équations suivantes :

$$(1) \qquad w(x) = y,$$
$$(2) \qquad {}^t w(x^*) = y^*$$

admettent une solution et une seule.

b) Ou bien les noyaux de w et de ${}^t w$ ne sont pas réduits à $\{0\}$ et, dans ce cas,

$$\dim \operatorname{Ker}(w) = \dim \operatorname{Ker}({}^t w).$$

L'équation (1) admet une solution si et seulement si y est orthogonal à $\operatorname{Ker}({}^t w)$, et l'équation (2) admet une solution si et seulement si y^* est orthogonal à $\operatorname{Ker}(w)$.

Cette alternative ne subsiste pas lorsque w est un endomorphisme continu d'un espace vectoriel localement convexe séparé E, même si E est un espace hilbertien et si w est hermitien. Néanmoins, à l'aide du théorème de finitude et de la théorie de Riesz, on démontre l'énoncé suivant :

Soit E un espace vectoriel localement convexe séparé, E' son dual topologique, u un automorphisme de E, v un endomorphisme compact de E et $w = u + v$. Alors l'alternative énoncée plus haut reste valable, *mutatis mutandis*. En particulier, les assertions suivantes sont équivalentes :

a) l'endomorphisme w est un automorphisme de E ;

b) l'endomorphisme w est injectif ;

c) l'endomorphisme w est surjectif ;

d) l'endomorphisme ${}^t w$ est un automorphisme de E' ;

e) l'endomorphisme ${}^t w$ est surjectif ;

f) l'endomorphisme ${}^t w$ est injectif.

Grâce aux résultats précédents, on peut préciser la nature de la *résolvante d'un endomorphisme compact*. Soit u un endomorphisme d'un espace vectoriel localement convexe séparé E. On dit qu'un nombre complexe A est une valeur *singulière* de u si A n'est pas seul et si $1/\lambda$ appartient à $\operatorname{sp}(u)$. Le complémentaire de l'ensemble des valeurs singulières est noté $\operatorname{reg}(u)$. Lorsque u est compact, $\operatorname{reg}(u)$ est un ouvert de C dont le complémentaire est discret. Considérons alors la résolvante R de u, c'est-à-dire l'application :

$$\mu \mapsto (\mathrm{I}_E - \mu u)^{-1}$$

Cette résolvante est holomorphe sur reg (u), et chaque valeur singulière A de u est un pôle d'ordre égal à la multiplicité *m(A)* de la valeur propre $1/\lambda$. Plus précisément, soit p_λ et q_λ les projecteurs **sur** F_ν et F',, où $\nu = 1/\lambda$. Alors u se décompose de la manière suivante :

$$u = r_\lambda + s_\lambda,$$

où $r_\lambda = q_\lambda u$ et $s_\lambda = p_\lambda u$. De plus :

a) L'endomorphisme r_λ est compact et λ est une valeur régulière de r_λ.

b) L'endomorphisme s_λ est continu de rang fini et A est l'unique valeur singulière de s_λ.

c) La résolvante de u est égale au produit des résolvantes de r_λ et de s_λ. Celle de r_λ est holomorphe au voisinage de A ; celle de s_λ admet A pour unique pôle. d'ordre *m(A).*

Cas des espaces hilbertiens

Soit E un espace hilbertien et u un endomorphisme continu de E. Supposons que u est normal, c'est-à-dire que $u^*u = uu^*$

(l'importance de cette classe d'endomorphismes provient du fait que les endomorphismes hermitiens, antihermitiens ou unitaires sont normaux). Alors la norme de u est égale au rayon spectral de u, c'est-a-dire au rayon du plus petit disque de centre 0 contenant le spectre de u. En particulier, un endomorphisme normal dont le spectre est réduit à $\{0\}$ est nul. En outre, toute valeur propre λ de u est d'indice 1, et l'on a :

$$E_{\bar{\lambda}}(u^*) = E_{\lambda}(u).$$

Enfin, les sous-espaces propres de u sont orthogonaux deux à deux. Néanmoins, il peut arriver qu'un endomorphisme continu normal (ou même hermitien) n'admette aucune valeur propre. Il en est ainsi de l'endomorphisme de multiplication par une fonction continue non constante à valeurs réelles dans l'espace hilbertien $L^2([a, b])$ des classes de fonctions de carré intégrable sur $[a, b]$. Toutefois, lorsque u est à la fois compact et normal, la théorie-de Riesz prend la forme achevée que voici :

Théorème spectral. Soit u un endomorphisme compact et normal d'un espace hilbertien E.

a) Le spectre de u est une partie compacte de C dont tout point autre que 0 est isolé.

b) Toute valeur spectrale non nulle de u est valeur propre de u.

c) Pour tout élément non nul λ de $\mathrm{sp}(u)$, le sous-espace propre E_λ est de dimension finie.

d) L'espace hilbertien E est somme hilbertienne des sous-espaces propres de u, c'est-à-dire que ces sous-espaces propres sont orthogonaux deux à deux et que leur somme directe est dense dans E. En particulier, **tout** vecteur x de E s'écrit d'une manière et d'une seule sous la forme :

$$x = \sum_{\lambda \in \mathrm{sp}(u)} x_\lambda,$$

où, pour tout élément λ de $\mathrm{sp}(u)$, x_λ appartient à E_λ. Dans ces conditions, on a les formules :

$$u(x) = \sum_{\lambda \in \mathrm{sp}(u)} \lambda x_\lambda, \qquad u^*(x) = \sum_{\lambda \in \mathrm{sp}(u)} \bar{\lambda} x_\lambda,$$

dites *formules de décomposition spectrale.* Il existe donc une base hilbertienne de E constituée de vecteurs propres de u.

On notera que le théorème précédent s'applique au cas d'un endomorphisme normal u d'un espace vectoriel hermitien de dimension finie. Dans ce cas, le spectre de u est fini, et les familles sommables intervenant dans les formules de décomposition spectrale se réduisent à des sommes finies. Il existe alors une base orthonormale de E constituée de vecteurs propres de u. En particulier, tout endomorphisme normal de E est diagonalisable. Par suite, pour toute matrice normale M, il existe une matrice diagonale D et une matrice unitaire P telles que :

$$M = PDP^{-1}$$

Ce résultat s'applique a la réduction des formes hermitiennes par rapport à une forme hermitienne positive non dégénérée (cf. FORMES QUADRATIQUES).

La théorie précédente s'applique aussi au cas des endomorphismes de puissance p-ième nucléaire : étant donné un espace hilbertien E et un endomorphisme hermitien positif h, le nombre :

$$\sum_{i \in I} (h(e_i) | e_i)$$

est indépendant du choix d'une base hilbertienne (e,), $i \in$ 1, de E. Ce nombre s'appelle *trace* de h et se note $\mathrm{tr}(h)$. Soit maintenant p un nombre réel supérieur à 1, E et F deux espaces hilbertiens et u un élément de C(E, F). On appelle valeur absolue de u l'endomorphisme $|u| = (u^*u)^{1/2}$; on dit que u est de puissance p-ième nucléaire si la trace de $|u|^p$ est finie. On pose alors :

$$\mathrm{N}p(u) = [\mathrm{tr}(|u|p)]^{1/p}.$$

Lorsque p = 1, de telles applications sont dites *nucléaires*, *ou* encore *traçables*. Lorsque p = 2, on les appelle *upplicutions de Hilbert-Schmidt*, *ou applications de carré traçable* (cf. équations DIFFÉRENTIELLES, chap. 3, et équations INTÉGRALES, chap. 5). Les applications de puissance p-ième nucléaire sont compactes ; plus précisément, soit u une application compacte de E dans F, soit (e,), $i \in$ I, une base hilbertienne de E constituée de vecteurs propres de $|u|$, et (μ_i), $i \in$ I, la famille correspondante de valeurs propres. Pour que u soit de puissancep-ième nucléaire, il faut et il suffit que :

$$\sum_{i \in \mathrm{I}} \mu_i^p < +\infty.$$

Dans ces conditions, on a :

$$\mathrm{N}p(u) = \left[\sum_{i \in \mathrm{I}} \mu_i^p\right]^{1/p}.$$

Les applications de puissance p-ième nucléaire constituent un sous-espace vectoriel, noté $\mathcal{L}^p$(E, F), de l'espace vectoriel des applications compactes, et l'application $u \mapsto$ N,,(u) est une norme sur $\mathcal{L}^p$(E, F), qui en fait un espace **de** Banach. Les applications de rang fini sont denses dans $\mathcal{L}^p$(E, F), et la convergence au sens de la norme N implique la convergence dans C(E, F). On peut alors développer une théorie des espaces $\mathcal{L}^p$(E, F) en tous points analogues à celle des espaces L^p intervenant dans la théorie de l'intégration (cf. INTÉGRATION ET MESURE, chap. 4).

Soit maintenant u un élément de $\mathcal{L}^1$(E) et (λ_n) la suite des valeurs propres non nulles de u, chacune d'elles étant écrite un nombre de fois égal à sa multiplicité. La série de terme général (A,,) est absolument convergente, et donc convergente ; sa somme s'appelle trace de u. De même, pour tout nombre complexe z, le produit infini de terme général $(1 + z\lambda_n)$ est convergent ; on pose :

$$\det(\mathrm{I}_{\mathrm{E}} + zu) = \prod_{n=0} (1 + z\lambda_n)$$

La fonction entière $z \mapsto \det(\mathrm{I}_{\mathrm{E}} + zu)$ s'appelle *déterminunt de Fredholm* de I'endomorphisme u ; on peut calculer son développement en série entière en introduisant les puissances extérieures hilbertiennes de u, notées $\Lambda^r(u)$:

$$\det(\mathrm{I}_{\mathrm{E}} + zu) = \sum_{r=0}^{+\infty} \mathrm{tr}(\Lambda^r(u))z^r.$$

On généralise ainsi la théorie du polynôme caractéristique, ce qui constitue la *théorie de Fredholm.*

3. Théorie spectrale de Hilbert

Soit u un endomorphisme continu normal d'un espace hilbertien E. La sous-algèbre unitaire fermée autoadjointe A de C(E) engendrée par u est une C*-algèbre commutative unitaire, dont le spectre s'identifie canoniquement à celui de u. De plus, la

transformation de Gelfand est un isomorphisme de A sur l'algèbre $\mathcal{C}(\mathrm{sp}(A))$ des fonctions continues sur le spectre de A. L'isomorphisme réciproque définit un morphisme φ de $\mathcal{C}(\mathrm{sp}(A))$ dans l'algèbre unitaire C(E) ; c'est l'unique morphisme de $\mathcal{C}(\mathrm{sp}(A))$ dans C(E) tel que $\varphi(z) = u$, où z est l'injection canonique de sp (A) dans C. Pour tout élément f de $\mathcal{C}(\mathrm{sp}\,(A))$, l'endomorphisme $\varphi(f)$ se note encore $f\,(u)$. Cette théorie permet donc de définir un *calcul fonctionnel* portant sur les fonctions continues de u. En particulier, $u^* = \varphi(\bar{z})$. L'objet de la théorie spectrale de Hilbert est d'étendre le calcul fonctionnel à des fonctions plus générales. On observe à cet effet que, pour tout couple (x, y) d'éléments de E, l'application :

$$f \mapsto \mu_{x,y}(f) = (f(u)(x)\,|\,y)$$

est (cf. INTÉGRATION ET MESURE, chap. 4) une *mesure de Radon* sur $\mathrm{sp}(u)$. De plus, l'application :

$$(x, y) \mapsto \mu_{x,y}$$

est sesquilinéaire hermitienne. Enfin, on a :

$$\|\mu_{x,y}\| \leqslant \|x\|.\|y\|,$$

Les mesures $\mu_{x,y}$ s'appellent *mesures spectrales* associées à u. On dit qu'une fonction f définie sur $\mathrm{sp}(u)$ à valeurs complexes est u-mesurable si, pour tout couple (x, y) d'éléments de E, cette fonction est $\mu_{x,y}$-mesurable. On note L"(u) l'algèbre des classes de fonctions u-mesurables essentiellement bornées et, pour tout nombre réel $p \geqslant 1$, on note $L^p(u)$ l'espace vectoriel des classes de fonctions u-mesurables appartenant à $L^p(\mu_{x,y})$ pour tout couple (x, y) d'éléments de E. On démontre alors le théorème fondamental suivant : Pour tout élément f de L'(u), il existe un élément et un seul de C(E), noté $f\,(u)$, tel que, pour tout couple (x, y) d'éléments de E, on ait :

$$\int_{\mathrm{sp}(u)} f(z)\mu_{x,y}(z) = (f(u)(x)\,|\,y).$$

C'est pourquoi *f (u) se* note encore :

$$\int_{\mathrm{sp}(u)} f(z)\mu(z),$$

où μ désigne la mesure vectorielle correspondant aux mesures scalaires $\mu_{x,y}$. En particulier, on a les formules de décomposition spectrale suivantes :

$$\int_{\mathrm{sp}(u)} z\mu(z), \quad u^* = \int_{\mathrm{sp}(u)} \bar{z}\mu(z).$$

De plus, l'application $f \mapsto f\,(u)$ est linéaire, et $[f(u)]^* = \bar{f}(u^*)$. En outre, pour toute suite (f_n) d'éléments de $L^1(u)$ convergeant simplement vers f, dominée par une fonction positive g appartenant à L'(u), la fonction f appartient à L'(u) et les endomorphismes *f,(u)* convergent fortement vers *f (u)* ; c'est le *théorème de convergence dominée de Lebesgue.* Enfin, pour **tout** élément f de $L^1(u)$ et pour tout élément g de $L^\infty(u)$, on a l'égalité :

$$(fg)(u) = f(u)g(u).$$

En particulier, l'application $u \mapsto f\,(u)$ définit un morphisme de la C*-algèbre L"(u) dans C(E), prolongeant ainsi le calcul fonctionnel aux éléments de L"(u) ; les éléments *f(u),* où f appartient à $L^1(u)$, appartiennent au bicommutant de A. Lorsque f est la fonction caractéristique d'une partie M de $\mathrm{sp}(u)$, $f(\chi_M)$ est un projecteur hermitien de E, noté P_M, appartenant au bicommutant de A. Les projecteurs P_M s'appellent *projecteurs spectraux* de u, et leurs images s'appellent *variétés spectrales* de u. Lorsque E est de dimension finie et que M est réduite à un point, on retrouve

la notion de sous-espace propre. Le calcul fonctionnel précédent permet de généraliser aux endomorphismes normaux la plupart des résultats de la théorie spectrale classique. On peut même définir un calcul fonctionnel portant sur les opérateurs normaux non bornés ; ses applications sont nombreuses (mécanique quantique, problèmes de Sturm-Liouville).

LUCIEN CHAMBADAL et JEAN-LOUIS OVAERT

Bibliographie

N. BOURBAKI, *Espaces vectoriels topologiques*, Masson, Parla, 1981 / R. DAUTRAY & J.-L. LIONS, *Analyse mathématique* **et** *calcul numérique* **pour** *les sciences et les techniques*, t. V : *Spectre des opérateurs*, *ibid.*, 1988 / J. DIEUDONNÉ, *Éléments d'analyse*, t. I et II, Gauthier-Villars, Paris. 3' éd. 1979-1983 / N. DUNFORD & J. SCHWARTZ, *Linear Operators*, 3 vol.. Wiley, 1988 / M. REED & B. SIMON, *Methods of Modern Mathematical Physics*, *4 vol.*, Acad. Press. New York, 1972-1979 / L. SCHWARTZ, *Analyse hilbertienne*, Hermann, 1979 / C. SCOVARNEC. *Algèbre spectrale*, 2 vol.. Publisud. Paris, 1990 / K. YOSIDA, *Functional Analysis*, Springer, New York, 6' éd. *1988.*

SYMBOLIQUE CALCUL

Le calcul symbolique est né au XIXe siècle d'une succession de démarches heuristiques et il a été particulièrement développé par Heaviside pour l'étude des circuits électriques.

Si l'on désigne par p la dérivation, p^2 désignera naturellement la double dérivation, $1/p$ l'intégration (encore faut-il choisir convenablement la « constante d'intégration »). L'opérateur qui à la fonction $f(t)$ fait correspondre la fonction $f(t - a)$ pourra, compte tenu de la formule de Taylor (cf. CALCUL INFINITÉSIMAL - Calcul à une variable, chap. 3), être représenté par e^{-ap}. En fait tous les opérateurs représentés ainsi symboliquement ont la propriété de permuter avec les translations dans le temps. Physiquement, cela signifie que ces opérateurs sont liés à des organes linéaires invariants dans le temps : si on décale dans le temps l'action exercée sur un tel organe, sa réponse subit le même décalage. Dans la terminologie moderne, ce sont des *opérateurs de convolution.* Par exemple, la dérivation est la convolution par la dérivée de la mesure de Dirac.

Le calcul symbolique a été justifié sur le plan théorique grâce à l'utilisation de la *transformation de Laplace.* Celle-ci associe à une fonction à support positif une fonction d'une variable complexe p. Un opérateur de convolution se transforme en un opérateur de multiplication par une fonction F de la variable complexe p. Enfin, grâce à la théorie des distributions, cette fonction F peut elle-même être considérée comme la transformée de Laplace de l'élément par lequel se fait la convolution. La transformation de Laplace opérant sur des éléments (fonctions ou distributions) à support positif, c'est à l'étude des régimes transitoires que le calcul symbolique est utilisé. Pour les systèmes à temps discret. une forme analogue de calcul symbolique a été développée sous le nom de *transformation en z.* Parmi les aspects qui ne pourront pas être traités ici, citons l'application aux systèmes différentiels à coefficients variables, l'application à la résolution de certaines équations aux dérivées partielles, la transformation de Laplace à plusieurs variables et les aspects numériques de l'utilisation de la transformation de Laplace.

Transformation de Laplace des fonctions et des mesures

Soit f une fonction à valeurs réelles ou complexes définie sur l'ensemble R des nombres réels et nulle pour les valeurs strictement négatives de la variable (c'est-à-dire que f est une fonction « à support positif »). Sa transformée de Laplace est la fonction $\mathfrak{L}f$ de la variable complexe p définie par la formule :

$$\mathfrak{L}f(p) = \int_0^{+\infty} f(t)\, e^{-pt}\, dt.$$

De même si μ est une mesure (cf. INTÉGRATION ET MESURE) sur R à support positif, c'est-à-dire telle que $\mu(\varphi) = 0$ pour toute fonction φ nulle pour les valeurs positives de la variable, sa transformée de Laplace est la fonction $\mathfrak{L}\mu$ de la variable complexe p définie par la formule :

$$\mathfrak{L}\mu(p) = \int e^{-pt}\, \mu(dt)\,;$$

si μ est une mesure de densitéfpar rapport à la mesure de Lebesgue, alors on a $\mathfrak{L}\mu = \mathfrak{L}f$. On notera par la suite Y la fonction définie par Y(t) = 1 si $t \geqslant 0$ et $Y(t) = 0$ si $t < 0$. On voit par un calcul élémentaire que la transformée de Laplace de Y est $1/p$. Plus généralement, la transformation de Laplace de f(t) $=$ Y($t)e^{\lambda t}$ est $1/(p-\lambda)$.

Au lieu du symbole $\mathfrak{L}$ pour représenter la transformation de Laplace, on utilise souvent un symbole, par exemple 7, pour relier les expressions analytiques d'une fonction et de sa transformée de Laplace. On écrira par exemple :

$$Y(t)\, e^{\lambda t} \sqsupset \frac{1}{p-\lambda},$$

et dans de nombreux ouvrages on sous-entend le facteur Y(t).

En fait, la transformée de Laplace d'une mesure μ n'est définie que pour les valeurs dep pour lesquelles la fonction e ^{pt} est intégrable par rapport à μ. On a le résultat suivant, facile à démontrer.

Théorème 1. Il existe un nombre ξ_0 tel que la fonction e^{-pt} soit intégrable par rapport à μ pour Rep $> \xi_0$ et non intégrable pour Rep $< \xi_0$, en désignant par Re p la partie réelle de p.

Le nombre ξ_0 est appelé *abscisse d'intégrabilité.* II peut être égal à $+\infty$ ou à $-\infty$. Pour une fonction f, on supposera f intégrable sur tout intervalle fini, de sorte que f est la densité d'une mesure μ. L'abscisse d'intégralibilité de $\mathfrak{L}\mu$ est appelée *abscisse de convergence absolue* de $\mathfrak{L}f$. Si Re $p > \xi_0$, l'intégrale :

$$\int_0^{+\infty} e^{-pt} f(t)\, dt$$

est absolument convergente, et, si Rep $< \xi_0$, cette intégrale n'est pas absolument convergente (elle peut être divergente ou « semi-convergente »). Par exemple, si $f(t) = Y(t)e^{\lambda t}$, on a ξ_0 = Re A. En fait, on a trouvé que :

$$\mathfrak{L}f(p) = \frac{1}{p-\lambda}\,;$$

c'est une fonction holomorphe pour Rep > Re λ, et elle se prolonge de façon naturelle au plan complexe. D'une façon plus générale, on a le résultat fondamental suivant.

Théorème 2. La transformée de Laplace d'une mesure μ est une fonction holomorphe pour Rep $> \xi_0$ (abscisse d'intégrabilité). La dérivée k-ième de $\mathfrak{L}\mu$ est donnée par :

$$(\mathfrak{L}\mu)^{(k)}(p) = (-1)^k [\mathfrak{L}(t^k \mu)](p).$$

Ainsi, avec f(t) 7 F(p), on a :

$$(-t)^k f(t) \; 7 \; F^{(k)}(p).$$

En particulier, si μ est à support compact, c'est-à-dire si toutes les fonctions continues sont μ-intégrables, alors sa transformée de Laplace est une fonction entière.

Théorème 3. Avec $\mu(dt)$ 7 F(p), on a :

$$e^{\lambda t}\mu(dt) \sqsupset F(p - \lambda),$$

autrement dit :

$$[\mathfrak{L}(e^{\lambda t}\mu)](p) = \mathfrak{L}\mu(p-\lambda).$$

Théorèmes 4, dits théortmes de la valeur finale et de la valeur initiale. Soit f une fonction à support positif ayant pour transformée de Laplace F, on a les résultats suivants :

a) Si f a une limite à droite pour t -- 0, alors on a :

$$\lim_{t \to +0} f(t) = \lim_{\substack{p \in \mathbf{R} \\ p \to +\infty}} pF(p).$$

b) Si f a une limite pour $t \to +\infty$, alors l'abscisse de convergence absolue ξ_0 est négative ou nulle, et l'on a :

$$\lim_{t \to +0} f(t) = \lim_{\substack{p \in \mathbf{R} \\ p \to 0}} pF(p).$$

Avant de donner la propriété principale de la transformée de Laplace des mesures, rappelons la définition suivante. Si μ et ν sont deux mesures sur R à support positif, alors :

$$\mu * \nu : \varphi \mapsto \int\int \varphi(x+y)\mu(dx)\nu(dy)$$

est une mesure sur R à support positif, appelée le produit de convolution des mesures μ et ν (c'est un cas particulier du produit de convolution de deux distributions ; cf. DISTRIBUTIONS, chap. 3). Si μ et ν sont de densités f et g par rapport à la mesure de Lebesgue, alors $\mu * \nu$ est de densité $f * g$, où :

$$(f*g)(t) = \int f(t-u)g(u)\,du$$

est le produit de convolution des fonctions f et g.

Théorème 5. Si $\mu * \nu$ désigne le *produit de composition* de deux mesures μ et v à support positif, on a :

$$\mathfrak{L}(\mu * \nu) = \mathfrak{L}\mu\mathfrak{L}\nu$$

pour les valeurs de la variable dont la partie réelle est supérieure aux deux abscisses d'intégrabilité.

Soit alors f une fonction à support positif, continûment dérivable pour $t > 0$, continue à droite pour $t = 0$. On a :

$$f(t) = \begin{cases} f(0) + \int_0^t f'(\theta)\,d\theta, & t \geqslant 0, \\ 0, & t < 0; \end{cases}$$

ce qui s'écrit aussi :

$$f = \mathrm{Y}f(0) + \mathrm{Y} * f',$$

d'où :

$$\mathfrak{L}f(p) = \frac{1}{p}f(0) + \frac{1}{p}\mathfrak{L}f',$$

c'est-à-dire :

$$\mathfrak{L}f'(p) = p\mathfrak{L}f(p) - f(0).$$

Si l'on a $f(0) = 0$, on trouve la formule simplifiée :

$$\mathfrak{L}f'(p) = p\mathfrak{L}f(p);$$

ce sont les discontinuités pour $t = 0$ qui compliquent les formules liant les transformées de Laplace d'une fonction à celles de ses dérivées. Or, la dérivation au sens **des** distributions fait intervenir ces discontinuités et permet, par suite, de conserver à ces formules leur forme la plus simple. Si

D désigne la dérivation au sens des distributions (cf. **DISTRIBUTIONS,** chap. 3), on aura :

$$\mathrm{D}f = f' + f(0)\delta,$$

δ étant la distribution de Dirac, et. par suite :

$$\mathcal{L}(\mathrm{D}f)(p) = p\,\mathcal{L}f(p).$$

Transformation de Laplace des distributions

Soit T une distribution à support positif telle que T = $\mathrm{D}^k f$, où f est une fonction pour laquelle $\mathcal{L}f$ a une abscisse de convergence absolue $\xi_0 \neq +\infty$. On posera :

$$\mathcal{L}\mathrm{T}(p) = p^k\,\mathcal{L}f(p)$$

pour Re$p > \xi_0$.

On montre aisément la cohérence de cette définition et la compatibilité avcc les définitions antérieures. On obtient en particulier $\mathcal{L}\delta^{(k)}(p) = p^k$. La généralisation aux distributions possédant la propriété indiquée des règles obtenues pour les fonctions et les mesures est aisée et donne les résultats suivants :

a) La transformée de Laplace d'une distribution T est holomorphe dans un demi-plan Rep > α, et l'on a :

(CT)' = — $\mathcal{L}(t\mathrm{T})$.

b) On a :

$$\mathcal{L}(\mathrm{T} * \mathrm{U}) = \mathcal{L}\mathrm{T}\;\mathcal{L}\mathrm{U};$$

en particulier si l'on fait **U =** $\delta^{(k)}$, on obtient :

$$\mathcal{L}(\mathrm{D}^k\mathrm{T})(p) = p^k\,\mathcal{L}\mathrm{T}(p).$$

c) On a :

$\mathcal{L}(e^{\lambda t}\mathrm{T})(p)$ **= CT@-h).**

T	CT	T	CT
δ	1	δ'_a	$e^{-ap}p$
δ_a	e^{-ap}	$\delta^{(k)}$	p^k
δ'	p	Y	$1/p$

T		CT
Y(t) ×	$e^{\lambda t}$	$\dfrac{1}{p-\lambda}$
	$\dfrac{1}{\omega}\sin \omega t$	$\dfrac{1}{p^2+\omega^2}$
	$\cos \omega t$	$\dfrac{p}{p^2+\omega^2}$
	$\dfrac{t^{\alpha-1}}{\Gamma(\alpha)}$ (a > 0)	$\dfrac{1}{p^\alpha}$
	$\dfrac{1}{\sqrt{t}}$	$\dfrac{\sqrt{\pi}}{\sqrt{p}}$
	$\dfrac{e^{\lambda t}t^{\alpha-1}}{\Gamma(\alpha)}$ (a > 0)	$\dfrac{1}{(p-\lambda)^\alpha}$

tabl. 1 - *Transformées* de Laplace *les plus usuelles*

Le tableau 1 donne les transformées de Laplace les plus usuelles.

Relations entre la transformation de Fourier et la transformation de Laplace

Dans ce chapitre, nous utiliserons la formule suivante pour définir la *transformation de Fourier* d'une fonction f (cf. analyse **HARMONIQUE,** chap. 3) :

$$(\mathcal{F}f)(y) = \int_{-\infty}^{+\infty} f(x)\,e^{-ixy}\,dx.$$

De la formule :

$$(\mathcal{L}f)(p) = \int_0^{+\infty} f(t)\,e^{-pt}\,dt,$$

supposée valable pour Re $p > \xi_0$ (abscisse de convergence absolue), il résulte que :

$$(\mathcal{L}f)(\xi + i\eta) = \int_0^{+\infty} f(t)\,e^{-\xi t}\,e^{i\eta t}\,dt = [\mathcal{F}(f(t)\,e^{-\xi t})](\eta).$$

Cette formule se généralise sans difficulté aux distributions. Avec $T = D^k f$, où f admet une transformée de Laplace définie pour $\operatorname{Re} p > \xi_0$, on aura, pour tout $\xi > \xi_0$,

$$\mathcal{L}T(\xi + i\eta) = [\mathcal{F}(T\, e^{-\xi t})](\eta).$$

En particulier, si $\xi_0 < 0$, on aura :

$$\mathcal{L}T(i\eta) = \mathcal{F}T(\eta).$$

La transformée de Laplace apparaît donc comme une extension convenable au plan complexe de la transformée de Fourier qui, elle, est une fonction de variable réelle. 11 convient de rappeler que cette extension n'a pu se faire que moyennant l'hypothèse que l'élément auquel on applique la transformation de Laplace était à support positif. Étant donné que la transformation de Fourier est injective, il en sera de même de la transformation de Laplace. On peut même reconstituer une fonction à partir de sa transformée de Laplace supposée connue sur une verticale du plan complexe d'abscisse $\xi > \xi_0$, abscisse de convergence absolue. On aura en général, compte tenu de la formule de réciprocité de la transformation de Fourier,

$$f(t) = \frac{1}{2i\pi}\int_{\xi-i\infty}^{\xi+i\infty} F(p)\, e^{pt}\, dp,$$

où l'intégrale figurant au second membre est prise dans le plan complexe sur la verticale d'abscisse ξ. L'holomorphie de F permet de modifier le chemin d'intégration. On peut également rechercher des conditions suffisantes assurant qu'une fonction $F(p)$ soit la transformée de Laplace d'une distribution. On a alors le résultat simple suivant.

Théorème. Si $F(p)$ est holomorphe pour $\operatorname{Re} p > c$ et vérifie $|F(p)| \leqslant C|p|^m$, alors F est la transformée de Laplace d'une distribution. Si l'on peut prendre $m = -2$, alors F est la transformée de Laplace d'une fonction.

Applications de la transformation de Laplace

L'application la plus répandue de la transformation de Laplace est la résolution des *équations de convolution*, et en particulier des équations différentielles linéaires à coefficients constants. Soit l'équation de convolution $a * x = b$, où a, b et x sont des fonctions à support positif. Si a, b, sont des transformées de Laplace A, B, X, on aura :

$$A(p)X(p) = B(p),$$

c'est-à-dire :

$$X(p) = B(p)/A(p).$$

La résolution de l'équation de convolution se ramène donc à la résolution d'une équation algébrique et à la recherche d'un élément ayant une transformée de Laplace donnée. Il est intéressant de noter que, pour les distributions à support positif, la convolution n'a pas de diviseurs de zéro. Une équation de convolution sur $\mathbf{R}_+$ ne peut donc avoir qu'une solution. Si l'usage de la transformation de Laplace fournit une solution (c'est-à-dire si a et b ont des transformées de Laplace et si $B(p)/A(p)$ est la transformée de Laplace d'une distribution), celle-ci est l'unique solution de l'équation.

Exemple 1. Soit à résoudre l'équation différentielle :

$$\frac{d^2x}{dt^2} + x = e^t$$

avec les conditions initiales :

$$x(0) = 1, \qquad \frac{dx}{dt}(0) = 2.$$

Si l'on ne s'intéresse qu'aux valeurs de $x(t)$ pour $t \geqslant 0$, on peut aussi bien suppo-

ser $x(t) = 0$ pour $t < 0$, à condition naturellement de supposer que le second membre est remplacé par 0 pour $t < 0$. Les conditions initiales indiquent alors des discontinuités de $x(t)$ et de dx/dt pour $t = 0$; et, pour en tenir compte, il suffit d'introduire les dérivées au sens des distributions :

$$\mathrm{D}x = \frac{dx}{dt} + \delta,$$

$$\mathrm{D}^2x = \mathrm{D}\left(\frac{dx}{dt}\right) + \delta' = \frac{d^2x}{dt^2} + 2\,\delta + \delta'.$$

L'équation différentielle se récrit alors :

$$(\mathrm{D}^2x - 2\,\delta - \delta') + x = \mathrm{Y}(t)\,e^t,$$

c'est-à-dire :

$$\mathrm{D}^2x + x = \mathrm{Y}(t)\,e^t + 2\,\delta + \delta'.$$

Soit X la transformée de Laplace de x. On obtient :

$$(p^2 + 1)\mathrm{X}(p) = \frac{1}{p-1} + 2 + p,$$

d'où :

$$\mathrm{X}(p) = \frac{p^2+p-1}{(p^2+1)(p-1)} = \frac{1}{2(p-1)} + \frac{p/2}{p^2+1} + \frac{3/2}{p^2+1}$$

et :

$$x(t) = \frac{1}{2}e^t + \frac{1}{2}\cos t + \frac{3}{2}\sin t, \qquad t \geqslant 0.$$

Exemple 2. Soit à résoudre l'équation :

$$\int_0^t \sin(t-\theta)\,x(\theta)\,d\theta = t^2, \qquad t \geqslant 0,$$

avec x à support positif. C'est une équation de convolution $a * x = b$, avec $a(t) = \mathrm{Y}(t)\sin t$ et $b(t) = \mathrm{Y}(t)t^2$. En prenant les transformées de Laplace, on obtient :

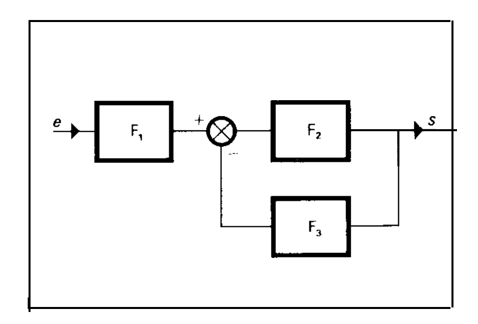
$$\frac{1}{p^2+1}\mathrm{X}(p) = \frac{2}{p^3},$$

d'où l'on déduit :

$$\mathrm{X}(p) = \frac{2}{p} + \frac{2}{p^3}, \quad x(t) = \mathrm{Y}(t)(2 + t^2).$$

Exemple 3. En automatique, **tout** organe linéaire invariant dans le temps établit une relation de la forme $s = f * e$ entre l'entrée e et la sortie s. Pour des raisons physiques, f est à support positif. Si S, F, E sont les transformées de Laplace de s, f, e, alors on $\mathrm{S}(p) = \mathrm{F}(p)\mathrm{E}(p)$, et F est appelée la *fonction de transfert* de l'organe. Dans le cas d'un système constitué de différents organes reliés entre eux, on obtient facilement la fonction de transfert F du système à partir de celles F_1, F_2, des différents organes. Par exemple, pour le système représenté par la figure, on a :

$$(\mathrm{F}_1(p)\mathrm{E}(p) - \mathrm{F}_3(p)\mathrm{S}(p))\,\mathrm{F}_2(p) = \mathrm{S}(p),$$



d'où :

$$\mathrm{F}(p) = \frac{\mathrm{S}(p)}{\mathrm{E}(p)} = \frac{\mathrm{F}_1(p)\,\mathrm{F}_2(p)}{1 + \mathrm{F}_3(p)\,\mathrm{F}_2(p)}.$$

Transformation en z

Soit a une suite réeiie ou complexe définie sur l'ensemble N des entiers positifs ou nuls. On appelle transformée en z de cette suite la fonction de variable complexe :

$$\mathrm{A}(z) = \sum_{n=0}^{\infty} \frac{a(n)}{z^n}$$

Il existe R $\in [0, +\infty]$ tel que cette série soit absolument convergente pour $|z| > R$. Le produit de composition $c = a * b$ de deux suites a et b sur N est défini par :

$$c(n) = \sum_{p+q=n} a(p)b(q),$$

et les transformées en z de a, b, c sont liées par la relation $C(z) = A(z)B(z)$. En particulier, l'opérateur dit opérateur d'avancement E, qui associe à toute suite a la suite Eu telle que $Eu(n) = a(n+1)$, est un opérateur de convolution :

$$Ea = a * \delta_1,$$

où δ_k est la suite égale à 1 pour $n = k$ et nulle ailleurs.

En posant $b = Eu$, on a la relation entre les transformées en z de a et b :

$$B(z) = \frac{1}{z} A(z).$$

Le tableau 2 donne quelques transformées en z de suites simples. L'utilisation

$a(n)$	$A(z)$
δ_k	$1/z^k$
λ^n	$z/(z-\lambda)$
n	$z/(z-1)^2$
n^2	$z(z+1)/(z-1)^3$
$\sin n\omega$	$z \sin\omega/(z^2 - 2z\cos\omega + 1)$
$\cos n\omega$	$(z^2 - z\cos\omega)/(z^2 - 2z\cos\omega + 1)$

tabl. 2 - *Quelques transformées en z de suites simples*

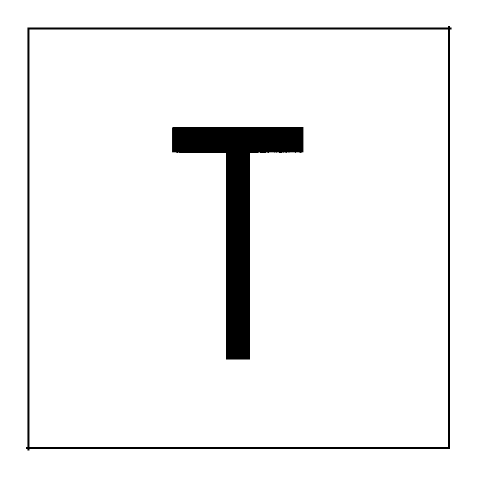
pratique de la transformation en z suppose de disposer de tables beaucoup plus importantes.

L'usage de la transformation en z permet la résolution des équations de récurrence linéaires à coefficients constants et plus généralement des équations de convolution sur N. Pour l'inversion de la transformation en z, on utilise généralement l'intégration dans le plan complexe. La transformation en z est largement utilisée pour l'étude des systèmes asservis à temps discret (souvent dénommés « échantillonnés »).

ROBERT PALLU DE LA BARRIÈRE

Bibliographie

R. BELLMAN, R. KALABA & J. LOCKETT, *Numerical Inversion of the Laplace Transform*, American Elsevier Publ. Co., New York, 1966 / O. HEAVISIDE, *Electromagnetic Theory*, 3ᵉ éd., Chelsea-New York, 1971 / J. HLADICK, *La Transformation de Laplace à plusieurs variables. Résolution des équations différentielles, intégrales et aux dérivées partielles*, Masson, Paris, 1969 / J. LAVOINE, *Calcul symbolique*, C.N.R.S., Paris, 1960 / W. R. LEPAGE, *Complex Variables and the Laplace Transform for Engineers*, McGraw-Hill, New York, 1961 / R. PALLU DE LA BARRIÈRE, *Cours d'automatique théorique*, Dunod, Paris, 1966 / J. R. RAGAZZINI & G. F. FRANKLIN, *Les Systèmes asservis échantillonnés* (*Sampled Data Control Systems*, 1958), traduit par un groupe de travail de la section genevoise de l'Association suisse pour l'automatique (Asspa), *ibid.*, Paris, 1962 / L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1951 / D. V. WIDDER, *The Laplace Transform*, Princeton Univ. Press, Princeton (N. J.), 1941 / L. A. ZADEH & C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.



TENSORIEL CALCUL

Introduit en 1900 par G. Ricci-Curbastro et T. Levi-Civita, le calcul tensoriel est un puissant outil de l'analyse mathématique ; très utile en mécanique classique, il est indispensable en mécanique relativiste.

Dans le présent article, E est une variété différentiable de dimension n. Rappelons rapidement ce que cela signifie. Au voisinage de chaque point m_0 de E, on peut trouver un système de coordonnées locales, c'est-à-dire repérer chaque point m par ses n coordonnées $u^1(m), \dots, u^n(m)$. Mais, au voisinage de m_0, il existe une infinité de systèmes de coordonnées raisonnables ; aucun d'eux ne joue, a priori, un rôle particulier. Notons $v^1(m), \dots, v^n(m)$ les coordonnées de m pour un autre système ; il existe n fonctions indéfiniment dérivables $\psi^1, \dots, \psi^n$, définies sur un ouvert de E_n, telles que, pour tout point m voisin de m_0 et pour tout j, $1 \leqslant j \leqslant n$, on ait :

$$v^j(m) = \psi^j(u^1(m), \dots, u^n(m)).$$

Pour étudier la variété E, on choisit un système de coordonnées dans lequel on va faire les calculs ; dans les formules que l'on écrira, on distinguera d'une part ce qui dépend du système de coordonnées choisi, et qui est en général dépourvu d'intérêt, d'autre part ce qui décrit des phénomènes intrinsèques. Parmi les plus importants des objets que l'on peut associer à une variété se trouvent les *tenseurs*. On se propose ici de les décrire dans un système de coordonnées et de voir comment cette description varie si l'on change de système. On appliquera ensuite ces calculs à l'étude des variétés pseudo-riemanniennes, en donnant des formules explicites pour la dérivée covariante.



1. Champs de vecteurs et formes de degré 1

Soit **un** système de coordonnées $u^1, \dots, u^n$ au voisinage du point m_0 de E, de coordonnées $u_0^1, \dots, u_0^n$. Toute fonction numérique f définie sur E au voisinage de m_0 apparaît comme une fonction f^u de n variables réelles : le nombre $f^u(u^1, \dots, u^n)$ est la valeur def au point de coordonnées $u^1, \dots, u^n$. On dit que f est de classe $\mathcal{C}^\infty$ si f^u est une fonction de classe $\mathcal{C}^\infty$ d'un ouvert de E_n dans $\mathbf{R}$. Cette définition semble privilégier un système de coordonnées, mais il n'en est rien : en utilisant le fait que les changements de coordonnées sont de classe $\mathcal{C}^\infty$, on démontre que, pour tout autre système de coordonnées $v^1, \dots, v^n$ qui à f associe la fonction de n variables

f^v, la fonction f^u est de classe $\mathcal{C}^\infty$ si et seulement si f^v est de classe $\mathcal{C}^\infty$.

Un champ de vecteurs sur E (ou, plus précisément, un champ de vecteurs de classe $\mathcal{C}^\infty$; cf. équations aux DÉRIVÉS PARTIELLES, chap. 1) est une correspondance X qui à toute fonction f de classe $\mathcal{C}^\infty$ associe une fonction $X(f)$ de classe $\mathcal{C}^\infty$ et vérifie les trois conditions suivantes :

a) Si f et g coïncident au voisinage d'un point m, alors $X(f)$ et $X(g)$ coïncident au voisinage de m ;

b) $X(f+\lambda g) = X(f) + \lambda X(g)$, où λ est une constante ;

c) $X(fg) = X(f)g + f\,X(g)$.

On définit sur l'ensemble des champs de vecteurs une structure de module sur l'anneau des fonctions de classe $\mathcal{C}^\infty$ (cf. algèbre LINÉAIRE ET MULTILINÉAIRE, chap. 9) en posant :

$$(X + \varphi Y)(f) = X(f) + \varphi Y(f)$$

pour toute fonction f. En associant à toute fonction f la fonction $\partial f^u/\partial u^i$, on définit un champ de vecteurs, que l'on notera $\partial/\partial u^i$. On montre que les n champs ainsi obtenus forment une base du module des champs ; c'est-à-dire que, pour tout champ de vecteurs X, il existe une famille $X^1, ..., X^n$ de fonctions de classe $\mathcal{C}^\infty$, et une seule, telle que :

$$(1) \qquad X = \sum_{i=1}^{n} X^i \frac{\partial}{\partial u^i};$$

les X^i sont appelées les coordonnées du champ X dans le système de coordonnées $(u^1, ..., u^n)$.

Une forme de degré 1 est une correspondance ω qui à tout champ de vecteurs X associe une fonction $\omega(X)$ de classe $\mathcal{C}^\infty$ et vérifie la condition suivante :

d) Pour tout couple (X, Y) de champs de vecteurs et toute fonction φ de classe $\mathcal{C}^\infty$, on a :

$$\omega(X + \varphi Y) = \omega(X) + \varphi\omega(Y).$$

Une forme de degré 1 est donc une application linéaire du module des champs dans l'anneau des fonctions de classe $\mathcal{C}^\infty$. En posant :

$$(\omega + \varphi\omega')(Y) = \omega(Y) + \varphi\omega'(Y)$$

pour tout champ de vecteurs Y, on munit l'ensemble des formes de degré 1 d'une structure de module sur l'anneau des fonctions de classe $\mathcal{C}^\infty$. La correspondance $X \mapsto X(f)$ est une forme de degré 1 pour toute fonction f de classe $\mathcal{C}^\infty$; on l'appelle la différentielle de f, et on la note df. Soit u^i la fonction qui à tout point m associe sa i-ième coordonnée $u^i(m)$ dans le système $(u^1, ..., u^n)$; les formes $du^1, ..., du^n$ constituent une base du module des formes de degré 1, c'est-à-dire que, pour toute forme ω, il existe une famille $\omega_1, ..., \omega_n$ de fonctions de classe $\mathcal{C}^\infty$, et une seule, telle que :

$$(2) \qquad \omega = \sum_{i=1}^{n} \omega_i du^i.$$

Les fonctions ω_i sont appelées les coordonnées de la forme ω dans le système $(u^1, ..., u^n)$. On vérifie que :

$$(3) \qquad du^i\left(\frac{\partial}{\partial u^k}\right) = \begin{cases} 0, & i \neq k, \\ 1, & i = k. \end{cases}$$

Il en résulte que la valeur de la forme :

$$\omega = \sum_{i=1}^{n} \omega_i du^i$$

sur le vecteur :

$$X = \sum_{i=1}^{n} X^i \frac{J}{\partial u^i}$$

est :

$$(4) \qquad \omega(X) = \sum_{i=1}^{n} \omega_i X^i.$$

Changement de coordonnées

Soit maintenant un nouveau système de coordonnées $(v^1, \dots, v^n)$. On a des fonctions de changement de coordonnées φ^i et ψ^j telles que, pour tout m, on ait respectivement :

$$u^i(m) = \varphi^i(v^1(m), \dots, v^n(m)),$$
$$v^j(m) = \psi^j(u^1(m), \dots, u^n(m)).$$

Posons :

$$\frac{\partial \psi^j}{\partial u^i}(u^1(m), \dots, u^n(m)) = \alpha_i^j(m),$$
$$\frac{\partial \varphi^i}{\partial v^j}(v^1(m), \dots, v^n(m)) = \beta_j^i(m).$$

Les matrices :

$$\alpha(m) = (\alpha_i^j(m)), \qquad \beta(m) = (\beta_j^i(m))$$

sont inversibles et inverses l'une de l'autre ; on démontre les formules :

$$(5) \begin{cases} \dfrac{\partial}{\partial v^j} = \displaystyle\sum_{i=1}^{n} \beta_j^i \frac{\partial}{\partial u^i}; & \dfrac{\partial}{\partial u^i} = \displaystyle\sum_{j=1}^{n} \alpha_i^j \frac{\partial}{\partial v^j} \\ du^i = \displaystyle\sum_{i=1}^{n} \beta_j^i \, dv^j; & dv^j = \displaystyle\sum_{i=1}^{n} \alpha_i^j \, du^i. \end{cases}$$

De ces formules on déduit que, si ω est une forme dc degré 1 ct X un champ de vecteurs dont les coordonnées dans les systèmes $(u^1, \dots, u^n)$ et $(v^1, \dots, v^n)$ sont respectivement les fonctions X^i, X'^j, ω_i et ω'_j, on a :

$$(6) \begin{cases} X'^j = \displaystyle\sum_{i=1}^{n} \alpha_i^j X^i; & X^i = \displaystyle\sum_{j=1}^{n} \beta_j^i X'^j \\ \omega'_j = \displaystyle\sum_{i=1}^{n} \beta_j^i \omega_i; & \omega_i = \displaystyle\sum_{j=1}^{n} \alpha_i^j \omega'_j. \end{cases}$$

2. Tenseurs

Tenseurs covariants

Un tenseur covariant à p variables τ est une fonction. définie sur l'ensemble des systèmes de p champs de vecteurs et à valeurs dans l'anneau des fonctions de classe $\mathcal{C}^\infty$, qui est linéaire par rapport à chacune de ses variables. c'est-à-dire qui vérifie, pour tout k tel que $1 \leqslant k \leqslant p$, la condition L_k suivante : Pour tout système de vecteurs $(X_1, \dots, X_p)$, avec $X_k = Y + \varphi Z$ où φ est une fonction de classe $\mathcal{C}^\infty$, on a :

$$\begin{aligned} \tau(X_1, \dots, X_p) &= \tau(X_1, \dots, X_{k-1}, Y, X_{k+1}, \dots, X_p) \\ &\quad + \varphi\tau(X_1, \dots, X_{k-1}, Z, X_{k+1}, \dots, X_p) \end{aligned}$$

Soit $\omega_1, \dots, \omega_p$ un système de p formes de degré 1. En associant à tout système de champs $(X_1, \dots, X_p)$ le produit :

$$\omega_1(X_1)\omega_2(X_2) \dots \omega_p(X_p),$$

on définit un tenseur covariant à p variables que l'on note :

$$\omega_1 \otimes \dots \otimes \omega_p.$$

En particulier. si l'on s'est donné un système de coordonnées $(u^1, \dots, u^n)$, alors. pour tout système $i = (i_1, \dots, i_p)$ d'indices compris entre 1 ct n, on a un tenseur :

$$du^{i_1} \otimes \dots \otimes du^{i_p};$$

on montre que les n^p tenseurs ainsi définis forment une base du module des tenseurs covariants à p variables ; c'est-à-dire que, pour tout tenseur covariant à p variables τ, il existe une famille $(\tau_{i_1, \dots, i_p})$ de n^p fonctions de classe $\mathcal{C}^\infty$, et une seule, telle que :

(7) $$\tau = \sum_i \tau_{i_1, \dots, i_p} du^{i_1} \otimes \dots \otimes du^{i_p} ;$$

les fonctions $\tau_{i_1, \dots, i_p}$ sont les coordonnées du tenseur dans le système de coordonnées choisi.

Si l'on se donne un autre système de coordonnées $(v^1, \dots, v^n)$, les formules (5) entraînent les formules (8) figurant dans le tableau.

On en déduit qu'entre les coordonnées $\tau_{i_1, \dots, i_p}$ du tenseur dans le système de coordonnées $(u^1, \dots, u^n)$ et ses coordonnées $\tau'_{j_1, \dots, j_p}$ dans le système de coordonnées $(v^1, \dots, v^n)$ on a les relations (9) du tableau.

Tenseurs contravariants

De la même façon, on appelle tenseur contravariant à p variables une application, linéaire par rapport à chaque variable, de l'ensemble des systèmes de p formes de degré 1 dans l'anneau des fonctions de classe $\mathcal{C}^\infty$. Pour tout système d'indices $i = (i_1, \dots, i_p)$, on définit le tenseur :

$$\frac{\partial}{\partial u^{i_1}} \otimes \dots \otimes \frac{\partial}{\partial u^{i_p}}$$

par la formule :

$$\frac{\partial}{\partial u^{i_1}} \otimes \dots \otimes \frac{\partial}{\partial u^{i_p}} (\omega_1, \dots, \omega_p) = \omega_1\left(\frac{\partial}{\partial u^{i_1}}\right) \dots \omega_p\left(\frac{\partial}{\partial u^{i_p}}\right).$$

8	$du^{i_1} \otimes \dots \otimes du^{i_p} = \sum_{1 \leqslant j_1 \leqslant n} \dots \sum_{1 \leqslant j_p \leqslant n} \beta^{i_1}_{j_1} \dots \beta^{i_p}_{j_p} dv^{j_1} \otimes \dots \otimes dv^{j_p}$ $dv^{j_1} \otimes \dots \otimes dv^{j_p} = \sum_{1 \leqslant i_1 \leqslant n} \dots \sum_{1 \leqslant i_p \leqslant n} \alpha^{j_1}_{i_1} \dots \alpha^{j_p}_{i_p} du^{i_1} \otimes \dots \otimes du^{i_p}$
9	$\tau_{i_1, \dots, i_p} = \sum_{1 \leqslant j_1 \leqslant n} \dots \sum_{1 \leqslant j_p \leqslant n} \alpha^{j_1}_{i_1} \dots \alpha^{j_p}_{i_p} \tau'_{j_1, \dots, j_p}$ $\tau'_{j_1, \dots, j_p} = \sum_{1 \leqslant i_1 \leqslant n} \dots \sum_{1 \leqslant i_p \leqslant n} \beta^{i_1}_{j_1} \dots \beta^{i_p}_{j_p} \tau_{i_1, \dots, i_p}$
10	$\tau = \sum_{1 \leqslant i_1 \leqslant n} \dots \sum_{1 \leqslant i_p \leqslant n} \tau^{i_1, \dots, i_p} \frac{\partial}{\partial u^{i_1}} \otimes \dots \otimes \frac{\partial}{\partial u^{i_p}}$
10′	$\tau^{i_1, \dots, i_p} = \sum_{1 \leqslant j_1 \leqslant n} \dots \sum_{1 \leqslant j_p \leqslant n} \beta^{i_1}_{j_1} \dots \beta^{i_p}_{j_p} \tau'^{j_1, \dots, j_p}$ $\tau'^{j_1, \dots, j_p} = \sum_{1 \leqslant i_1 \leqslant n} \dots \sum_{1 \leqslant i_p \leqslant n} \alpha^{j_1}_{i_1} \dots \alpha^{j_p}_{i_p} \tau^{i_1, \dots, i_p}$

Les n^p tenseurs ainsi définis forment une base du module des tenseurs contravariants à p variables ; donc, pour tout tenseur contravariant à p variables τ, on a une famille $(\tau^{i_1 \dots i_p})$ de fonctions de classe $\mathcal{C}^\infty$, et une seule, telle que la relation (10) du tableau soit satisfaite : les fonctions $\tau^{i_1 \dots i_p}$ sont les coordonnées du tenseur dans le système de coordonnées $(u^1, \dots, u^n)$.

Si maintenant les fonctions $\tau'^{j_1 \dots j_p}$ sont les coordonnées de τ dans le système $(v^1, \dots, v^n)$, on a les relations (10') du tableau.

Tenseurs de variance mixte

Soit (I, J) une partition de l'ensemble des entiers compris entre 1 et p. Un tenseur τ de variance (I, J) est une fonction de p variables $(A_1, \dots, A_p)$, linéaire par rapport à chacune d'elles, où A_α est une forme de degré 1 si $\alpha \in I$ et un champ de vecteurs si $\alpha \in J$. L'ensemble I est l'ensemble des indices contravariants de τ et J l'ensemble des indices covariants. Chaque système de coordonnées $(u^1, \dots, u^n)$ définit une base (ayant n^p éléments) du module des tenseurs de variance (I, J) ; donc τ est déterminé, dans le système de coordonnées choisi, par n^p fonctions coordonnées.

Par exemple, pour $p = 3$, un tenseur τ de variance ({1, 3}, 2) associe à tout triplet (ω, X, π), où X est un champ de vecteurs et ω, π des formes de degré 1, une fonction $\tau(\omega, X, \pi)$. Pour tout triplet (i_1, i_2, i_3) d'indices compris entre 1 et n, en associant à (ω, X, π) la fonction produit :

$$\omega\left(\frac{\partial}{\partial u^{i_1}}\right) du^{i_2}(X)\, \pi\left(\frac{\partial}{\partial u^{i_3}}\right),$$

on définit un tenseur de variance ({1, 3}, 2), que l'on note :

$$\frac{\partial}{\partial u^{i_1}} \otimes du^{i_2} \otimes \frac{\partial}{\partial u^{i_3}};$$

les tenseurs de ce type forment une base, c'est-à-dire que, τ étant donné, il existe une famille et une seule de n^3 fonctions de classe $\mathcal{C}^\infty$, que l'on notera $(\tau^{i_1 i_3}_{i_2})$, telle que :

$$\tau = \sum_{i_1, i_2, i_3 = 1}^{n} \tau^{i_1 i_3}_{i_2} \frac{\partial}{\partial u^{i_1}} \otimes du^{i_2} \otimes \frac{\partial}{\partial u^{i_3}}.$$

Conventions de sommation

Dans tout ce qui précède, on a décrit un certain nombre de familles d'objets : bases du module des champs ou du module des formes de degré 1 ; coordonnées des champs, des formes, des tenseurs ; dérivées partielles des changements de coordonnées. Les objets de chacune de ces familles sont alors repérés par un ou plusieurs indices compris entre 1 et n. On a placé certains de ces indices en haut, d'autres *en* bas. A priori, le choix de la place de ces indices paraît assez arbitraire ; il n'en est rien. En effet, si on regarde toutes les formules de sommation qui ont été écrites, on constate que (à condition de considérer que, dans $\partial/\partial u^i$, l'indice i est en bas) :

a) Chaque fois qu'on a sommé par rapport à un indice, celui-ci apparaît deux fois dans la formule, une fois en haut et une fois en bas ;

b) Chaque fois qu'un indice apparaît deux fois, il est une fois en haut et une fois en bas, et on somme par rapport à lui.

Ces deux remarques permettent de retrouver assez facilement toutes les formules que l'on a écrites et, d'autre part, d'éliminer celles qui n'ont aucun sens intrinsèque. Par exemple, si on a deux champs X et Y de coordonnées X^i et Y^i dans un certain système, on n'écrira jamais :

$$\sum X^i Y^i,$$

car cette quantité dépend du système de coordonnées choisi ; mais, si ω est une forme de degré 1, de coordonnées ω_i, la somme :

$$\sum_i \omega_i X^i$$

est indépendante du système de coordonnées choisi.

On utilise cette convention pour l'écriture des tenseurs ; on écrit les indices contravariants en haut et les indices covariants en bas. C'est ce que l'on a fait dans les exemples cités plus haut.

Comme les sommations que l'on doit faire sont indiquées par les indices, on prend l'habitude de supprimer les signes Σ de sommation ; c'est la *convention d'Einstein*. Ainsi, les formules (9) s'écrivent :

$$\tau_{i_1,\ldots,i_p} = \alpha_{i_1}^{j_1} \ldots \alpha_{i_p}^{j_p} \tau'_{j_1,\ldots,j_p},$$
$$\tau'_{j_1,\ldots,j_p} = \beta_{j_1}^{i_1} \ldots \alpha_{j_p}^{i_p} \tau_{i_1,\ldots,i_p},$$

La sommation par rapport aux indices j dans la première formule et par rapport aux indices i dans la seconde est indiquée par le fait que chacun de ces indices se trouve une fois en bas et une fois en haut.

La contraction

Ce principe de sommation, énoncé pour des indices de deux tenseurs différents, s'étend aux indices d'un même tenseur. Considérons, par exemple, un tenseur τ à quatre indices, de coordonnées τ^j_{ikl} dans un certain système de coordonnées ; les quantités τ^i_{ikl}, c'est-à-dire :

$$\sum_{i=1}^{n} \tau^i_{ikl},$$

sont les coordonnées d'un tenseur à deux variables, appelé le contracté **de** τ par rapport **aux** deux premiers indices. De façon générale, chaque fois qu'un tenseur a un indice covariant et un indice contravariant, on peut les contracter et on obtient ainsi un tenseur qui a deux variables de moins que le tenseur que l'on contracte.

3. Variétés pseudo-riemanniennes

Nous écrirons désormais toutes les formules dans un système de coordonnées fixe $(u^1, \ldots, u^n)$, d'ailleurs arbitraire. Soit g un tenseur covariant à deux variables qui est symétrique, c'est-à-dire que $g(X, Y) = g(Y, X)$ quels que soient X et Y. Les valeurs en un point m des coordonnées g_{ij} de g forment une matrice carrée d'ordre n, et le déterminant de cette matrice dépend du système de coordonnées choisi ; mais, s'il est nul dans un système, il l'est dans tous les autres. Si ce déterminant ne s'annule en aucun point m, on dit que g est une structure pseudo-riemannienne sur E. Dans ce qui suit, on fixe une fois pour toutes une structure pseudo-riemannienne g.

Soit X un champ de vecteurs ; la correspondance qui à tout champ de vecteurs Y associe $g(X, Y)$ est une forme de degré 1 ; on la note $X^\flat$. La correspondance qui à X associe $X^\flat$ est un isomorphisme du module des champs sur le module des formes ; elle est donc biunivoque, et :

$$(X + \varphi Y)^\flat = X^\flat + \varphi Y^\flat.$$

Si on note X^i les coordonnées de X, et $X_i^\flat$ celles de $X^\flat$, on a, en utilisant les conventions de notations ci-dessus.

$$X_i^\flat = g_{ij} X^j. \qquad (11)$$

Cet isomorphisme **b** se prolonge au cas des tenseurs : Si τ est un tenseur dont l'indice a est contravariant, on définit un tenseur $\tau^{\flat(\alpha)}$ dont la variance diffère de celle de τ par le fait que a y est covariant

en posant, pour tout système de champs et de formes, (..., X, . ..). où l'on n'a pas écrit les variables d'indice différent de α :

$$\tau^{\flat(\alpha)}(..., X, . ..) = \tau(..., X^{\flat}, . ..),$$

les coordonnées $\tau^{\flat(\alpha)}{}_{..i\alpha}$ de $\tau^{\flat}(\alpha)$ proviennent des coordonnées $\tau^{\,i\alpha...}$ de τ, grâce à la formule :

$$(12) \qquad \tau^{\flat(\alpha)}{}_{...i_\alpha...} = g_{i_\alpha j_\alpha}\tau^{\,j_\alpha}$$

Pour tout m, la matrice des coefficients $g_{ij}(m)$ est inversible ; soit $g^{ij}(m)$ la matrice inverse. Les fonctions g^{ij} sont de classe C^∞. Avec les conventions de notation que l'on a faites, le fait que la matrice $(g_{ij}(m))$ soit l'inverse de la matrice $(g^{ij}(m))$ se traduit par la formule :

$$(13) \qquad g^{ij}g_{jk} = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases}$$

De cette relation résulte que :

$$g^{ij}g_{jk}g_{il} = g_{kl}\,;$$

autrement dit, le tenseur g est obtenu à partir du tenseur g^* de coordonnées g^{ij} par application de $\flat$ (1) et $\flat$ (2).

On note $\sharp$ les isomorphismes inverses des isomorphismes $\flat$ et on démontre que :

a) Pour toute forme ω de degré 1, on a :

$$(\omega^{\sharp})^k = g^{kj}\omega_j\,;$$

b) Pour tout tenseur τ dont l'indice α est covariant, on a :

$$(\tau^{\sharp(\alpha)})^{\,i_\alpha\cdot} = g^{i_\alpha j_\alpha}\tau_{..\,j_\alpha}$$

On voit que l'on peut, grâce aux isomorphismes $\flat$ et $\sharp$, modifier à volonté les variances des indices d'un tenseur. En particulier, on peut les rendre tous covariants ou tous contravariants. Les coordonnées du tenseur covariant associé à τ sont appelées les coordonnées covariantes de τ et celles du tenseur contravariant associé à τ sont appelées les coordonnées contravariantes de τ. Les g^{ij} sont donc les coordonnées contravariantes du tenseur métrique g et les g_{ij} ses coordonnées covariantes. De même les X^i sont les coordonnées contravariantes du champ X et les $X_i = g_{ij}X^j$ ses coordonnées covariantes.

4. La dérivée covariante

Soit D une connexion linéaire sur E : à tout couple (X, Y) de champs de vecteurs elle associe un champ de vecteurs $D_X Y$ tel que l'on ait les relations :

$$(\alpha) \qquad D_{X+\phi X'}Y = D_X Y + \phi D_{X'}Y,$$
$$(\beta) \qquad D_X(Y + \phi Y') = D_X Y + \phi D_X Y' + X(\phi)Y.$$

Soit X^i les coordonnées de X, soit Y^i celles de Y et Γ^i_{jk} celles de :

$$D_{\frac{\partial}{\partial u^j}}\frac{\partial}{\partial u^k}$$

dans le système de coordonnées $(u^1, ... u^n)$. Des conditions (α) et (β) on déduit :

$$(14) \qquad D_X Y = X^j\left(\frac{\partial Y^i}{\partial u^j} + \Gamma^i_{jk}Y^k\right)\frac{\partial}{\partial u^i}.$$

Si D est la dérivée covariante, on montre que les Γ^i_{jk} sont donnés par la formule :

$$(15) \qquad \Gamma^i_{jk} = \frac{1}{2}g^{il}\left(\frac{\partial g_{kl}}{\partial u^j} + \frac{\partial g_{lj}}{\partial u^k} - \frac{\partial g_{jk}}{\partial u^l}\right);$$

on les appelle les symboles de Christoffel. On remarquera que, contrairement à ce qu'une analogie de notations pourrait laisser penser, les symboles de Christoffel ne se comportent pas comme les coordonnées d'un tenseur ; c'est-à-dire que, si l'on exprime la dérivée covariante dans un autre système de coordonnées, on obtient

d'autres fonctions Γ'^{v}_{uw} qui ne sont pas égales à :

$$\Gamma^{j}_{ik}\,\alpha^{i}_{u}\,\beta^{v}_{j}\,\alpha^{k}_{w}.$$

Ces calculs permettent d'écrire de façon très simple le système différentiel des géodésiques : Soit γ une courbe paramétrée proportionnellement à sa longueur ; c'est une géodésique si et seulement si :

$$D_{d\gamma/dt}\frac{d\gamma}{dt} = 0,$$

c'est-à-dire si, en notant γ^i les fonctions coordonnées de γ, on a :

$$\frac{d^2\gamma^i}{dt^2} + \Gamma^{i}_{jk}\frac{d\gamma^j}{dt}\frac{d\gamma^k}{dt} = 0, \tag{16}$$

quel que soit i.

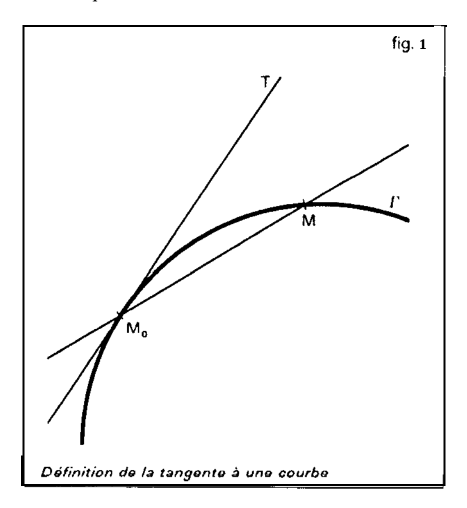
CLAUDE MORLET

Bibliographie

É. CARTAN, *Leçons sur la géométrie des espaces de Riemann*, éd. rev. et augm.. Gauthier-Villars, Paris, 1963 / J. C. H. GERRETSEN, *Lectures on Tensor Calculus and Differential Geometry*, Groningen, 1962 / S. GOLAB, *Tensor Calculus*, Elsevier Scientific Publ., Amsterdam-New York, 1974 / A. LICHNEROWICZ, *Éléments de calcul tensoriel*, Armand Colin, Paris. 1950 ; *Théorie globale des connexions et des groupes d'holonomie*, Masson. Paris-Rome, 1955 ; *Algèbre et analyse linéaire,* Dunod, Paris, 1970 / J. S. SOKOLNIKOFF, *Tensor Analysis. Theory and Applications to Geometry and Mechanics of Continua*, 2' éd., Wiley, New York-Londres-Sydney, 1964.

TOPOLOGIE GÉNÉRALE

Les notions de continuité et de limite ont une origine intuitive et l'on se propose d'analyser ici cette intuition. Considérons, par exemple, la description de la tangente T à une courbe (fig. 1) telle qu'on la trouve dans les manuels classiques de géométrie élémentaire : Si M varie sur Γ, la corde M_0M varie continûment et, si M tend vers M_0, la corde M_0M a une position limite qui est T.



Définition de la tangente à une courbe

En disant que M_0M varie continûment, on exprime que, si M s'approche indéfiniment d'un point M,, la droite M_0M s'approche indéfiniment de la droite M_0M_1 ; en disant que M_0M a une position limite T, on exprime que, si M s'approche indéfiniment de M_0, la droite M_0M s'approche indéfiniment de T. On peut donc donner les définitions suivantes :

- L'applicationfde X dans Y est continue en x_1 si une condition suffisante pour que f(x) soit voisin de $f(x_1)$ est que x soit assez voisin de x_1 ;
- L'application f de X x_0 dans Y a une limite y_0 en x_0, si une condition suffisante pour que $f(x)$ soit voisin de y_0 est que x soit assez voisin de x_0.

Pour que ces définitions deviennent des définitions mathématiques, il faut donner un sens précis aux termes « $f(x)$ voisin de $f(x_1)$ (ou de y_0) » et « x assez voisin de x_1

(ou de x_0) ». Dans les chapitres 1 et 2, on s'occupera d'abord de définir cette notion de *voisinage,* puis on donnera les principales propriétés des fonctions continues et des limites. Les chapitres 3 et 4 seront consacrés à l'étude de deux classes d'espaces topologiques très importantes, les *espaces compacts* et les *espaces connexes.*

La notion d'espace topologique contient en particulier celle d'espace métrique (cf. espaces MÉTRIQUES) dont l'étude est une excellente introduction à la topologie générale.

1. Espaces topologiques

Voisinages et continuité

On a vu que, pour définir les notions de limite et de continuité. on devait donner un moyen de savoir si deux points sont voisins (resp. assez voisins). Pour cela, il est assez naturel de mesurer la distance de ces deux points. On peut donc parler de continuité ou de limites pour les applications de X dans Y, si l'on a défini la distance entre les points de X et la distance entre les points de Y, c'est-à-dire si X et Y sont des espaces métriques (cf. espaces MÉTRIQUES).

Ce point de vue est suffisant tant que Xet Y sont $\mathbf{R}$, $\mathbf{R}^p$, les surfaces de $\mathbf{R}^3$, etc., et, plus généralement, pour tous les problèmes géométriques. C'est l'analyse qui a mis ses lacunes en évidence ; il arrive, en effet, que l'on dispose d'applications de $\mathbf{R}^p$ dans un ensemble de fonctions E qui, pour des raisons propres au problème à résoudre, doivent être considérées comme continues, mais qu'il n'existe aucune métrique sur E qui les rende continues. Il faut donc donner un moyen, autre que la distance, pour savoir si deux éléments a et b de E sont voisins.

Pour cela, on se donne une famille $\mathcal{V}_a$ **de** sous-ensembles de E que l'on appelle les *voisinages* de a dans E ; pour dire « comment b est voisin de a », on dit dans quel voisinage de a il se trouve. Si, pour tout élément a de E, on a défini les voisinages de a dans E, on dit que E est un *espace topologique* ; les éléments de E sont alors appelés des *points.* Si E et F sont deux espaces topologiques et sifest une application de E dans F, on dit quefest *continue* au point a de E si : Pour tout voisinage V de $f(a)$ dans F, il existe un voisinage W de a dans E tel que, pour **tout** point b de W, le point $f(b)$ soit dans V. On dit quefest continue si elle est continue en chaque point de E. Tout espace métrique X devient naturellement un espace topologique si l'on choisit pour voisinages d'un point x les sous-ensembles de X qui contiennent une boule de centre x et de rayon strictement positif. On vérifie alors **que,** si X et Y sont métriques, pour une application f de X dans Y les deux définitions de la continuité que l'on a données coïncident.

On impose aux voisinages de vérifier les quatre conditions suivantes :

(V,) Tout voisinage de a contient a ;

(V_2) Tout sous-ensemble de l'espace E qui contient un voisinage de a est un voisinage de a ;

(V,) L'intersection d'un nombre fini de voisinages de a est un voisinage de a ;

(V,) Pour tout voisinage V de a, il existe un voisinage W de a tel que V soit voisinage de chacun des points de W.

Ces conditions permettent d'étendre au cas des espaces topologiques les principales propriétés des *fonctions continues* de $\mathbf{R}$ dans R. En particulier, toute somme, tout produit et tout quotient de fonctions numériques continues sont encore des fonctions numériques continues. On démontre aussi que la composée de deux applications continues est une application continue.

Ouverts et fermés

On dit qu'un sous-ensemble U de l'espace topologique E est ouvert s'il est voisinage de chacun de ses points. Les ouverts d'un espace topologique E vérifient les trois propriétés suivantes :

(O_1) L'ensemble E et l'ensemble vide sont ouverts ;

(O_2) Toute réunion d'ouverts est un ouvert ;

(O_3) Toute intersection d'un nombre *fini* d'ouverts est un ouvert.

La structure topologique d'un espace est déterminée par la connaissance de ses ouverts ; en effet, les voisinages d'un point a de E sont les sous-ensembles de E qui contiennent un ouvert qui contient a.

Il est clair que la donnée des ouverts de E est équivalente à celle des sous-ensembles de E dont le complémentaire dans E est ouvert ; ces sous-ensembles sont appelés les *fermés* de E ; ils vérifient les trois conditions suivantes :

(F_1) L'ensemble E et l'ensemble vide sont fermés ;

(F_2) Toute intersection de fermés est un fermé ;

(F_3) Toute réunion d'un nombre *fini* de fermés est un fermé.

Soit A un sous-ensemble d'un espace topologique E. D'après (O_2), la réunion de **tous** les ouverts de E contenus dans A est un ouvert qui est évidemment le plus grand ouvert (au sens de l'inclusion) contenu dans A ; on l'appelle *l'intérieur* de A et on le note A. L'intérieur d'un ensemble peut être vide sans que cet ensemble le soit, comme on le voit en prenant, par exemple, pour E l'ensemble R des nombres réels muni de sa topologie usuelle, et pour A l'ensemble Q des nombres rationnels. Les ensembles ouverts sont caractérisés par le fait qu'ils sont égaux à leur intérieur.

De même, d'après (F_2), l'intersection de tous les fermés contenant A est un fermé A qui est le plus petit fermé contenant A : on l'appelle la *fermeture ou* l'*adhérence* de A. On vérifie facilement qu'un point $x \in E$ appartient à $\bar{A}$ si et seulement si, pour tout voisinage V de x, on a $V \cap A \neq \emptyset$; un tel point est dit *adhérent* à A. On dit enfin que A *est partout dense* dans E si $\bar{A} = E$, ce qui revient à dire que tout ouvert non vide de E rencontre A.

La considération des ouverts et des fermés donne une caractérisation très simple des applications continues. En effet. pour qu'une application f de E dans F soit continue, il **faut** et il suffit que, pour tout ouvert V de F, l'ensemble $f^{-1}(V)$ soit un ouvert de E ; ou encore, en termes de fermés : pour tout fermé A de F, l'ensemble $f^{-1}(A)$ est fermé dans E.

Exemples

On trouvera dans de nombreux articles du présent ouvrage des exemples d'ensembles munis d'une topologie, Voici quelques rappels et exemples complémentaires.

1. Sur tout ensemble X, on appelle *topologie grossière* la topologie dont les seuls ouverts sont X et l'ensemble vide. Tout point de X a alors un seul voisinage qui est X lui-même.

2. Il existe une topologie sur X pour laquelle tout sous-ensemble qui contient x est un voisinage de x ; c'est la *topologie discrète* ; tout sous-ensemble de X est alors à la fois ouvert et fermé.

3. Tout espace métrique X a une topologie naturelle (cf. espaces MÉTRIQUES). Les voisinages d'un point x sont les sous-ensembles de X, qui contiennent une boule de centre x dont le rayon est de la forme $1/n$, avec n entier. Pour qu'une topologie puisse être définie par une métrique, il est donc nécessaire que tout point x possède une

suite de voisinages $(V_x^1, \ldots, V_x^n, \ldots)$ telle que tout voisinage de x contienne l'un des V_x^n ; cette remarque permet de montrer que les topologies définies aux exemples (6) et (7) ci-dessous ne peuvent pas être déduites d'une métrique.

4. Si A est un sous-ensemble de l'espace topologique E, les sous-ensembles de A de la forme A ∩ U, où U est un ouvert de E, forment les ouverts d'une topologie sur A ; c'est la topologie induite sur A par la topologie de E ; muni de cette topologie, A est appelé un sous-espace *topologique* de E.

5. Si E et F sont deux espaces topologiques, on appelle pavé ouvert de E X F les sous-ensembles de la forme U X V, où U est un ouvert de E et V un ouvert de F ; les réunions de pavés ouverts sont les ouverts d'une topologie sur E X F, appelée *topologie produit.* Remarquons que la topologie naturelle de $\mathbf{R}^2$ est la topologie produit de la topologie de R par elle-même.

6. Soit E l'ensemble des applications de R dans R ; pour tout nombre x et tout ouvert U de R, notons V(x, U) l'ensemble des éléments g de E tels que g(x) appartienne à U. Parmi les topologies sur E dont les V(x, U) sont des ouverts, il en existe une qui a le moins d'ouverts ; c'est la *topologie de la convergence simple.* Ses ouverts sont les réunions d'intersections *finies* d'ensembles V $(x$, U).

7. Soit encore E l'ensemble des applications de R dans R ; pour Ω ouvert de $\mathbf{R}^2$, notons V(Ω) l'ensemble des éléments g de E tels que, pour tout x, le point $(x$, g(x)) soit dans Ω. Les V(n), lorsque Ω parcourt l'ensemble des ouverts de $\mathbf{R}^2$, sont les ouverts d'une topologie sur E, que l'on appelle la topologie $\mathscr{C}^0$. On voit facilement qu'une application φ d'un espace topologique A dans E, muni de cette topologie $\mathscr{C}^0$, est continue si et seulement si l'application :

@ : A X R - R,

qui à (a, t) associe la valeur de <p(a) en t, est continue, l'ensemble A X R étant muni de la topologie produit (cf. exemple 5).

Dans les exemples 6 et 7 précédents, on a défini deux topologies distinctes sur l'ensemble E des applications de R dans R. mais il en existe bien d'autres : topologie discrète ; topologie grossière ; topologie de la convergence uniforme, définie par la distance :

$$d(f, g) = \sup_{x \in \mathbf{R}} (\inf(|f(x) - g(x)|, 1)).$$

Sur un ensemble donné, il existe ainsi beaucoup de topologies ; pour certains ensembles, en particulier pour les ensembles de fonctions, plusieurs d'entre elles ont un réel intérêt ; mais, sur R et sur $\mathbf{R}^n$, on n'utilise pratiquement que l'une d'elles, celle qui est déduite de la distance euclidienne.

Homéomorphismes

Une application f de l'espace topologique X dans l'espace topologique Y est appelée un *homéomorphisme* si elle est bijective et si elle est continue ainsi que son inverse.

Il est important de noter qu'une application bijective et continue n'est pas nécessairement un homéomorphisme ; par exemple, si X est le sous-espace de R formé de]a, b[et des points a et c, avec $c > b$, l'applicationfde X dans $[a, b]$, définie par $f(t) = t$ si $t \neq c$ et f(c) = b, est bijective et continue, mais son inverse n'est pas continue.

Recollements de topologies

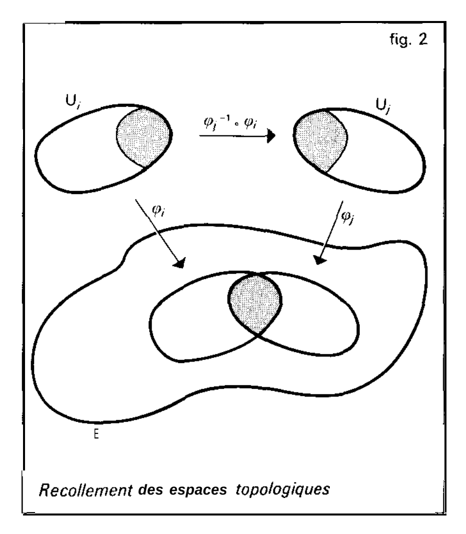
Soit (U,), $i \in I$, une famille d'espaces topologiques, et, pour tout i, une injection

φ_i de U, dans un ensemble E. On suppose vérifiées les trois conditions suivantes :

a) E est la réunion des images des φ_i ;

b) Pour tout couple (i, j), l'ensemble $\varphi_i^{-1}(\varphi_j$ (U,)) est un ouvert de U, ;

c) Pour tout couple (i, j), l'application $\varphi_j^{-1} \circ \varphi_i$ réalise un homéomorphisme de $\varphi_i^{-1}(\varphi_j$ (U,)) sur $\varphi^{-1}_i(\varphi_i(U_i))$ comme l'indique la figure 2.

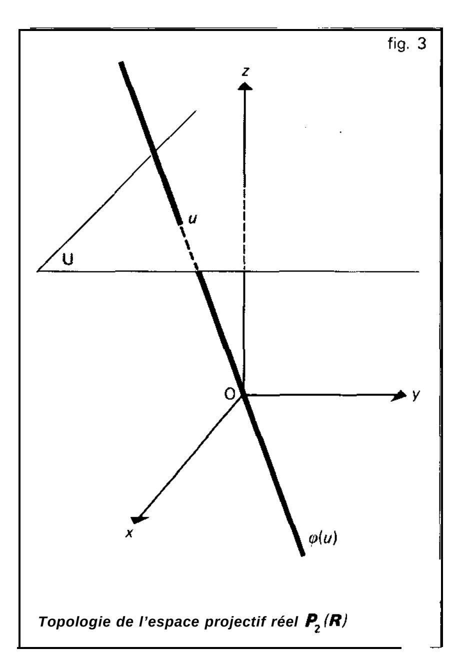


Recollement des espaces topologiques

Alors, il existe une topologie et une seule sur E, telle que, pour tout i, l'application φ_i soit un homéomorphisme de U_i sur le sous-espace φ_i (U,) de E. Un sous-ensemble Ω de E est ouvert pour cette topologie si et seulement si, pour tout i, $\varphi_i^{-1}(\Omega)$ est ouvert dans U_i.

Parmi les topologies définies par ce procédé de recollement, citons la topologie de l'espace projectif réel $\mathbf{P}_n(\mathbf{R})$, c'est-à-dire de l'ensemble des sous-espaces vectoriels de dimension 1 de $\mathbf{R}^{n+1}$: droites de $\mathbf{R}^{n+1}$ passant sur l'origine. Pour tout sous-espace affine U de $\mathbf{R}^{n+1}$ qui ne passe pas par l'origine, on définit une injection :

$$\varphi : \mathbf{U} \rightarrow \mathbf{P}_n(\mathbf{R})$$



Topologie de l'espace projectif réel $\boldsymbol{P}_2(\boldsymbol{R})$

en associant à $u \in$ U le sous-espace vectoriel $\varphi(u)$ engendré par u (fig. 3). Les conditions (a), (h) et (c) sont vérifiées. Si $n = 1$ ou 2, l'espace P,,(R) s'identifie à l'ensemble des droites du plan ou de l'espace de la géométrie élémentaire qui passent par un point M_0 donné ; c'est la topologie que nous venons de décrire qui est utilisée quand, dans la définition de la tangente, on parle de la limite d'une droite passant par M_0 (fig. 1).

On définit de façon analogue une topologie sur l'espace projectif complexe P,,(C), c'est-à-dire sur l'ensemble des sous-espaces vectoriels de dimension 1 de $\mathbf{C}^{n+1}$.

2. Limites

Exemples

Dans l'analyse classique, le mot limite peut désigner des choses apparemment très

diverses dont on va citer quelques exemples.

1. *Limite d'une suite numérique.* Soit (x_n) une suite de nombres ; on dit qu'elle converge et que sa limite est y si, quel que soit le nombre strictement positif ε, il existe un entier N tel que, pour tout $n \geqslant$ N, on ait l'inégalité y x_n $\leqslant \varepsilon$.

2. *Limite uniforme d'une suite de fonctions.* Soit (f_n) une suite de fonctions numériques définies sur un ensemble X ; on dit qu'elle converge uniformément et que sa limite est g si, quel que soit le nombre strictement positif ε, il existe un entier N tel que, pour tout $n \geqslant$ N et pour tout élément z de X, on ait l'inégalité $|g(z) - f_n(z)$ $\leqslant \varepsilon$.

3. *Limite en $+\infty$ d'une fonction numérique définie sur l'intervalle $[a, +\infty[$.* Soit f une fonction numérique définie sur l'intervalle $[a, +\infty[$; on dit qu'elle a une limite en $+\infty$ et que cette limite est le nombre y si, quel que soit le nombre strictement positif ε, il existe un nombre N tel que l'inégalité $x \geqslant \sup\ (a,$ N) entraîne l'inégalité $y - f(x)$ $\leqslant \varepsilon$.

4. *Limite à droite en a d'une fonction numérique définie sur $]a, b[$.* Soit f une fonction numérique définie sur $]a, b[$; on dit que fa une limite à droite en a et que cette limite est y si, quel que soit le nombre strictement positif ε, il existe un nombre strictement positif α tel que, pour tout point x qui vérifie $a < x \leqslant a + \alpha$, on ait $y - f(x)$ $\leqslant \varepsilon$.

Filtres

II existe une certaine parente entre ces définitions ; pour dégager ce qu'elles ont de commun, H. Cartan a introduit la notion de filtre. Un *filtre* sur l'ensemble A est, par définition, un ensemble $\mathcal{F}$ de parties de A, qui vérifie les trois conditions suivantes :

a) Tout élément de $\mathcal{F}$ est non vide ;

b) L'intersection d'un nombre fini d'éléments de $\mathcal{F}$ est un élément de $\mathcal{F}$;

c) Toute partie de A qui contient un élément de $\mathcal{F}$ est elle-même un élément de 3.

Si f est une application de A dans un espace topologique X, on dit que le point y de X est limite defsuivant le filtre $\mathcal{F}$ si, pour tout voisinage V de y dans X, il existe un élément F de $\mathcal{F}$ tel que *f(a)* soit dans V chaque fois que a appartient à F.

Montrons que l'on a obtenu le résultat cherché, c'est-à-dire que les quatre notions de limites définies plus haut sont des cas particuliers de limite suivant un filtre.

Sur l'ensemble N des entiers naturels, les complémentaires des parties finies forment un filtre $\mathcal{F}_{\mathrm{N}}$ (que l'on appelle souvent filtre de Fréchet). Dire que la suite numérique (x_n) converge et a pour limite y (exemple 1), c'est dire que y est limite de la fonction de N dans R qui à n associe x_n, suivant le filtre $\mathcal{F}_{\mathrm{N}}$.

Dire que la suite de fonctions numériques (f_n) converge uniformément et qu'elle a pour limite g (exemple 2), c'est dire que, si l'on note E l'ensemble des fonctions de X dans R et si l'on munit E de la distance de la convergence uniforme, g est limite suivant le filtre $\mathcal{F}_{\mathrm{N}}$ de l'application de N dans E qui à n associe f_n.

Dire que y est la limite de f en $+\infty$ (exemple 3), c'est dire que y est la limite de f suivant le filtre $\mathcal{F}_{+\infty}$ formé des complémentaires des parties majorées de $[a, +\infty[$.

Dire que y est la limite à droite en a de ia fonction numérique f définie sur $]a, b[$ (exemple 4), c'est dire que y est la limite de f suivant le filtre $\mathcal{F}_{a+}$ formé des intersections de $]a, b[$ et des voisinages de a dans R.

On démontre que, si f et g sont des applications de A dans R qui ont des

limites y et z suivant le filtre $\mathcal{F}$, alors $y + z$ (resp. yz) est limite de $f + g$ (resp. limite de fg) suivant le filtre $\mathcal{F}$.

Notons encore que, si A est un espace topologique, les voisinages d'un point a forment un filtre sur A ; dire que l'application f de A dans l'espace topologique X est continue en a, c'est dire que f a pour limite $f(a)$ suivant ce filtre.

Séparation

Une suite numérique ne peut avoir deux limites ; de même, une fonction numérique ne peut avoir deux limites en $+\infty$ (ou en a). Mais, si l'ensemble A est muni du filtre $\mathcal{F}$ et si f est une application de A dans un espace topologique X, deux points distincts de X peuvent être limites de f suivant le filtre $\mathcal{F}$; par exemple, si X est muni de la topologie grossière, tout point de X est limite de f suivant le filtre $\mathcal{F}$. Pour avoir l'unicité des limites pour les applications à valeurs dans X, on doit supposer que X vérifie la condition suivante : Deux points distincts de X possèdent des voisinages disjoints. On dit alors que X est *séparé*. Tous les espaces métriques sont séparés, à cause de la relation $d(x, y) = 0 \Rightarrow x = y$; c'est le fait que R est séparé qui permet de faire les raisonnements classiques dits « par passage à la limite » ou « par continuité ».

3. Espaces compacts

Les intervalles fermés bornés de R ont des propriétés topologiques remarquables, connues depuis très longtemps ; ces propriétés découlent toutes du fait qu'ils vérifient la condition suivante, appelée condition de Borel-Lebesgue (cf. le théorème (7) du chapitre 4 de l'article CALCUL INFINITÉSIMAL - Calcul à une variable).

Condition (BL). On dit que l'espace topologique E vérifie la condition de Borel-Lebesgue si, quelle que soit la famille d'ouverts (U_i), $i \in I$, de E telle que :

$$\bigcup_{i \in I} U_i = E,$$

il existe un sous-ensemble *fini* J de I tel que :

$$\bigcup_{i \in J} U_i = E.$$

Par définition, on dit qu'un espace topologique est *compact* s'il est séparé et s'il vérifie la condition de Borel-Lebesgue. Cette condition est équivalente à chacune des deux suivantes.

Condition (BL)'. Quelle que soit la famille (F_i), $i \in I$, de fermés de E d'*intersection vide*, il existe un sous-ensemble *fini* J de I tel que l'intersection des F_i, pour $i \in J$, soit vide.

Condition (BL)''. Quelle que soit la famille (F_i), $i \in I$, de fermés de E, si, pour tout sous-ensemble *fini* J de I,

$$\bigcap_{i \in J} F_i$$

est non vide, il existe (au moins) un point de E qui appartient à tous les F_i.

Exemples

L'intervalle $[a, b]$ de R est compact. Plus généralement, un sous-espace A de R est compact si et seulement s'il est fermé et borné. De la même façon, les sous-espaces compacts de $\mathbf{R}^n$ sont les fermés bornés. Tout sous-espace fermé d'un compact est compact. Tout produit d'espaces compacts est compact.

Propriétés

Citons les plus importantes propriétés des espaces compacts.

1. Si X est compact et Y séparé, l'image d'une application continue de X dans Y est un sous-espace compact de Y. En particulier, si A est compact, l'image d'une application continue f de A dans R est un sous-espace compact de R ; c'est donc un sous-ensemble fermé borné de R ; c'est pourquoi f est une application bornée et atteint ses bornes.

2. *Propriété de Bolzano-Weierstrass*. Soit (u_n), $n \in$ N, une suite de points du compact A. Alors, il existe un point a de A tel que tout voisinage de a contienne u_n pour une infinité de valeurs de n ; un tel point a est appelé une *valeur d'adhérence* de la suite.

La démonstration, par l'absurde, est la suivante : Si, pour tout point x de A, il existait un voisinage ouvert V_x de x qui ne contienne qu'un nombre fini des u_n, alors, comme N est infini, l'ensemble A ne pourrait pas être recouvert par un nombre fini des V_x.

Réciproquement, cette propriété de Bolzano-Weierstrass entraîne la propriété de Borel-Lebesgue si l'espace considéré est un espace métrique.

3. Tout espace métrique compact est complet (cf. espaces MÉTRIQUES, chap. 3).

4. Tout sous-espace compact d'un espace topologique E est un fermé de E.

5. Toute fonction continue d'un espace compact dans un espace métrique est uniformément continue (cf. espaces MÉTRIQUES, chap. 2).

Espaces localement compacts

On dit qu'un espace topologique est localement compact s'il est séparé et si chacun de ses points possède un voisinage compact. On se convainc de l'importance des espaces localement compacts en remarquant que tous les espaces de la géométrie ($\mathbf{R}$, $\mathbf{R}^n$, surfaces, courbes, variétés différentiables...) sont localement compacts. Tout espace localement compact est homéomorphe à un ouvert d'un espace compact. En fait, on peut même s'arranger pour que cet ouvert soit le complémentaire d'un unique point de l'espace compact (compactification d'Alexandrof).

Parmi les propriétés les plus utiles des espaces localement compacts, citons les deux suivantes, qui concernent les problèmes de prolongement de fonctions continues lorsque E est un espace localement compact « dénombrable à l'infini », c'est-à-dire tel qu'il existe une famille dénombrable (K_n), $n \in$ N, de sous-espaces compacts de E, avec :

$$\bigcup_{n \in \mathrm{N}} \mathrm{K}_n = \mathrm{E}.$$

1. *Théorème de Tietze.* Toute fonction continue d'un fermé de E dans R est la restriction d'une fonction continue définie sur E tout entier.

2. *Existence de «partitions de l'unité »*. Pour tout recouvrement de E par des ouverts (U_i), $i \in$ I, il existe une famille de fonctions (f_i), $i \in$ I, telle que :

a) quel que soit le point x n'appartenant pas à U_i, on a $f_i(x) = 0$,

b) en chaque point x, il n'existe qu'un nombre fini des f_i qui ne s'annulent pas,

c) pour tout point x, on a :

$$\sum_{i \in \mathrm{I}} f_i(x) = 1 ;$$

une telle famille de fonctions est appelée une partition de l'unité relative au recouvrement (U_i), $i \in$ I, donné.

4. Espaces connexes

On regardant une figure géométrique, chacun sait dire si elle est formée de plusieurs

morceaux disjoints. La connexité est la notion mathématique qui correspond à cette réalité physique. Si la figure F est formée de deux morceaux disjoints A et B, tout point de F assez voisin de A est encore dans A, et tout point de B assez voisin de B est encore dans B. Donc A et B sont des ouverts non vides et disjoints de F. Inversement, si F est d'un seul tenant, il n'existe pas de partition de F en deux ouverts non vides disjoints. Mathématiquement, on exprime ce fait en disant que F est connexe.

Les sous-espaces connexes de R sont les intervalles de R, ouverts, semi-ouverts ou fermés, bornés ou non. Si X est connexe et si $f : X \to Y$ est une application continue, l'image de f est un sous-espace connexe de Y. En particulier, l'image d'une application continue $f : [a, b] \to R$ est un intervalle de R. C'est le *théorème de la valeur intermédiaire*, qui s'énonce comme suit.

Théorème. Si f est une application continue de $[a, b]$ dans R, quel que soit z compris entre f(a) et $f(b)$, il existe un point c de $]a, b[$ tel que $f(c) = z$ (cf. CALCUL INFINITÉSIMAL - Calcul à une variable, chap. 9, théorème 14 *bis).*

Notons encore que, si X est connexe, toute fonction localement constante de X dans Y, c'est-à-dire telle que tout point x de X possède un voisinage sur lequel f est constante, est constante dans X tout entier. Le principe des zéros isolés pour les fonctions analytiques (cf. FONCTIONS ANALYTIQUES - Fonctions analytiques d'une variable complexe, chap. 1) utilise ce type d'argumentation.

CLAUDE MORLET

Bibliographie

G. CHOQUET, *Cours de topologie : espaces topologiques et espaces métriques, fonctions numériques,* Masson, Paris, 2ᵉ éd. rev. 1984 / J. DIEUDONNÉ, *Éléments d'analyse*, t. I : *Fondements de l'analyse moderne,* Gauthier-Villars, Paris, 2ᵉ éd. 1979 / L. SCHWARTZ, *Analyse*, vol. I : *Théorie des ensembles et topologie,* Hermann, Paris, 1991.

TOPOLOGIQUE ALGÈBRE

L'algèbre topologique est consacrée à l'étude d'ensembles munis d'une topologie et d'une structure algébrique définie par des lois de composition continues (cf. TOPOLOGIE GÉNÉRALE ; ALGÈBRE). Les exemples les plus importants sont les groupes topologiques, les espaces topologiques à groupe d'opérateurs, les espaces vectoriels topologiques (cf. ESPACES VECTORIELS TOPOLOGIQUES) et les anneaux topologiques dont les corps topologiques sont un cas particulier.



1. Groupes topologiques

Un groupe topologique est un espace topologique G muni d'une loi de composition interne *continue* : $G \times G \to G$ vérifiant les axiomes d'une loi de groupe, notée multiplicativement et telle que l'application $x \mapsto x^{-1}$ de G dans lui-même soit aussi continue (c'est alors un homéomorphisme involutif de G sur G). Un morphisme de groupes topologiques est une application continue qui est un homomorphisme de groupes.

Soit G un groupe topologique ; si $a \in G$, la translation à gauche :

$$\gamma_a : s \mapsto as, \quad s \in G,$$

est un homéomorphisme de G sur lui-même qui transforme l'élément neutre e en a (l'homéomorphisme réciproque est γ_a^{-1}). Ainsi les voisinages de a sont les images y,(V) = aV par y, des voisinages V de e (on pourrait aussi utiliser les translations à droite). Par suite, la loi de groupe de G étant supposée connue, sa topologie est déterminée par le filtre $\mathscr{V}$ des voisinages de e ; ce filtre a les propriétés suivantes :

(GV,) Quel que soit U $\in \mathscr{V}$, il existe V $\in \mathscr{V}$ tel que VV C U ; cette condition exprime que la loi de G est continue en (e, e).

(GV,) Quel que soit U $\in \mathscr{V}$, on a $U^{-1} \in$ 'Y; cette condition exprime que x $\mapsto x^{-1}$ est continue en e.

(GV,) Quels que soient U $\in \mathscr{V}$ et a $\in$ G, on a $aUa^{-1} \in \mathscr{V}$; autrement dit, l'automorphisme intérieur $x \mapsto axa^{-1}$ est un homéomorphisme de G sur G.

Notons que la condition (GV,) est automatiquement vérifiée si G est commutatif. Inversement, considérons un groupe G et un filtre $\mathscr{V}$ sur G possédant les propriétés précédentes ; il existe sur G une topologie et une seule faisant de G un groupe topologique et pour laquelle $\mathscr{V}$ est le filtre des voisinages de e.

Tout homomorphisme d'un groupe topologique dans un autre qui est continu en l'élément neutre est continu partout ; donc c'est un morphisme de groupes topologiques.

Exemples

1. On fait d'un groupe quelconque G un groupe topologique en le munissant de la topologie *discrète.*

2. La droite numérique réelle R, munie de la loi *additive,* est un groupe topologique. Il en est de même du groupe *multiplicatif* R* = R − {0} et de son sous-groupe $\mathbf{R}_+^*$ formé des nombres réels strictement positifs. L'application exponentielle x $\mapsto e^x$, avec $x \in$ R, est un isomorphisme de groupes topologiques de R sur $\mathbf{R}_+^*$; l'isomorphisme réciproque est donné par le logarithme (cf. EXPONENTIELLE ET LOGARITHME).

3. Plus généralement, on peut considérer les espaces numériques $\mathbf{R}^m$ avec la loi additive (et en particulier $\mathbf{C}^m = \mathbf{R}^{2m}$). Le groupe multiplicatif se généralise en le groupe $\mathbf{GL}(m,$ R) des matrices carrées inversibles d'ordre m ; c'est un ouvert de M,,,(R) = $\mathbf{R}m^2$ et on le munit de la topologie induite. De même $\mathbf{GL}(m,$ C) a une structure de groupe topologique ; pour $m = 1$, c'est le groupe multiplicatif C* (cf. nombres COMPLEXES).

4. D'une manière encore plus générale, tous les groupes de Lie sont des groupes topologiques ; rappelons que tout homéomorphisme continu d'un groupe de Lie dans un autre est automatiquement analytique, de sorte que, si **un** groupe topologique admet une structure analytique, celle-ci est unique (cf. GROUPES - Groupes de Lie).

5. Soit E un espace vectoriel normé, réel ou complexe (cf. espaces vectoriels NORMÉS) ; sa topologie et sa loi additive en font un groupe topologique. Plus généralement, on peut prendre pour E un espace vectoriel topologique (cf. espaces vectoriels TOPOLOGIQUES). Le groupe multiplicatif des éléments inversibles d'une algèbre de Banach est un groupe topologique (cf. algèbres NORMÉES).

6. Considérons un groupe G et un ensemble $\mathscr{S}$ de sous-groupes distingués de G qui est filtrant pour la relation $\supset$; le filtre engendré par $\mathscr{S}$ définit sur G une topologie qui en fait un groupe topologique. Par exemple, on peut prendre pour $\mathscr{S}$ l'ensemble des sous-groupes d'indice fini de G.

Un cas particulier important est celui *où* l'on se donne une *filtration* de G par des sous-groupes distingués, c'est-a-dire une suite décroissante (G_n), avec $n \in Z$, de sous-groupes distingués. Par exemple, la topologie p-adique sur le groupe additif Q est définie par la filtration $(p^n\mathbf{Z})$ oùp est un nombre premier (cf. théorie des NOMBRES - Nombres *p*-adiques).

7. Pour qu'un groupe topologique G soit *séparé*, il faut et il suffit que l'intersection des voisinages de l'élément neutre e soit réduite à $\{e\}$, ou encore que {e} soit *fermé.*

8. Soit G et H des groupes topologiques ; le produit G X H muni de la topologie produit et de la loi de groupe produit est un groupe topologique. On peut de même considérer le produit d'une famille quelconque de groupes topologiques. Par exemple, $\mathbf{R}^m$ est le produit de m facteurs identiques à R. Le groupe multiplicatif C* est isomorphe au produit $\mathbf{R}_+^*$ X U, où U désigne le groupe multiplicatif de nombres complexes de module 1 (muni de la topologie induite par C).

Sous-groupes, groupes quotients

Soit G un groupe topologique et H un sous-groupe de G ; la topologie induite sur H par celle de G en fait un groupe topologique. Par exemple, les groupes classiques, qui sont des sous-groupes de $\mathbf{GL}(m$, R) ou de $\mathbf{GL}(m$, C), ont une structure de groupe topologique : ce sont des groupes de Lie (cf. GROUPES - Groupes classiques et géométrie). L'adhérence $\overline{H}$ d'un sous-groupe H de G est encore un sous-groupe ; si H est distingué, il en est de même de $\overline{H}$. Pour qu'un sous-groupe soit ouvert, il faut et il suffit qu'il ait au moins un point intérieur, et alors il est aussi fermé ; ainsi, lorsque la topologie de G est définie par un ensemble filtrant de sous-groupes distingués (cf. exemple 6), ces sous-groupes sont à la fois ouverts et fermés et G est totalement discontinu. La composante connexe de l'élément neutre e dans un groupe topologique G est toujours un sous-groupe distingué fermé G_0 ; c'est le sous-groupe engendré par un voisinage arbitraire de e.

Les sous-groupes fermés de R distincts de R sont de la forme $a\mathbf{Z}$, avec $a \in$ R ; ils sont discrets (les autres sous-groupes sont partout denses). Plus généralement, les sous-groupes fermés de $\mathbf{R}^m$ sont de la forme :

$$V \times W \simeq \mathbf{R}^s \times \mathbf{Z}^t,$$

où V est un sous-espace vectoriel, isomorphe à $\mathbf{R}^s$, de $\mathbf{R}^m$, et $W \simeq \mathbf{Z}^t$ est un sous-groupe engendré par une partie d'une base d'un supplémentaire de V ; un tel sous-groupe est de rang $s + t$ et est discret dans le cas où $s = 0$. On appelle réseau de $\mathbf{R}^m$ un sous-groupe discret de rang maximum, c'est-à-dire engendré par une base de $\mathbf{R}^m$, donc isomorphe à $\mathbf{Z}^m$.

Parmi les groupes topologiques localement compacts, les groupes *de Lie* (cf. exemple 4) sont caractérisés par la propriété d'admettre un voisinage de l'élément neutre e ne contenant pas d'autre sous-groupe que $\{e\}$ (A. Gleason et H. Yamabe, 1953). On en déduit qu'un groupe localement compact et localement connexe qui est de dimension topologique finie est un groupe de Lie ; il en est ainsi, en particulier, des groupes topologiques qui sont des *variétés* topologiques, localement isomorphes à des ouverts de $\mathbf{R}^m$; cela fournit une réponse (partielle) au cinquième problème de Hilbert. À l'opposé, tout voisinage de e dans un groupe localement compact totalement discontinu contient un sous-groupe ouvert ; on en déduit que la

topologie d'un *groupe compact totalement discontinu* est définie par un ensemble filtrant de sous-groupes distingués (cf. exemple 6).

Soit H un sous-groupe distingué d'un groupe topologique G. La topologie et la loi de groupe quotient font de G/H un groupe topologique, qui est séparé dans le cas où H est fermé et qui est discret dans le cas où H est ouvert. Par exemple, si G est un groupe compact totalement discontinu, son quotient G/H par un sous-groupe distingué ouvert est discret et compact, donc fini. Le *tore* à une dimension $T = \mathbf{R}/\mathbf{Z}$ est un exemple important de groupe quotient ; il est compact, car il est séparé (**Z** étant fermé dans R) et image de l'intervalle compact [0, 1]. Les groupes quotients séparés de $\mathbf{R}^m$ sont de la forme $\mathbf{T}^t \times \mathbf{R}^{m-r}$, où $r = s + t$ est le rang du noyau ; un tel groupe est compact seulement si $r =$ m. Le quotient $\pi_0(G) = G/G_0$ d'un groupe topologique G par la composante connexe G_0 de son élément neutre est un groupe topologique séparé et totalement discontinu, discret si G est localement connexe.

Tout morphismef: G → H de groupes topologiques admet une décomposition canonique :

$$G \to G/\mathrm{Ker} f \xrightarrow{f} f(G) \hookrightarrow H,$$

où Kerfest le sous-groupe distinguéf '(e) de G ; le morphisme fest bijectif, mais ce n'est pas un homéomorphisme en général, On dit que f est un *morphisme strict* si f est un homéomorphisme ; cela a lieu en particulier si G/Ker f est compact (resp. localement compact dénombrable à l'infini) et $f(G)$ séparé (resp. de Baire). Par exemple, le morphisme :

$$x \mapsto e^{2\pi i x}, \quad x \in R,$$

de R dans C* est un morphisme strict :

$$\mathbf{R} \to \mathbf{T} \simeq \mathbf{U} \hookrightarrow \mathbf{C}^*.$$

Limites projectives de groupes topologiques

Soit I un ensemble ordonné filtrant supérieurement (c'est-à-dire tel que toute paire d'éléments ait une borne supérieure). Un *système projectif*, indexé par I, de groupes topologiques est un système (G_i, f_{ij}) où les G_i, pour $i \in I$, sont des groupes topologiques et où f_{ij}, pour $i, j \in 1$ avec $i \leqslant j$, est un morphisme de G_{jw} dans G_i. Par définition, la limite projective :

$$G = \varprojlim G_i$$

de ce système est le sous-groupe du groupe produit πG_i formé des éléments x = (x_i) tels que $f_i(x_j) = x_i$ pour tout $i \leqslant j$ et G a donc une structure de groupe topologique. Lorsque les morphismes canoniques G → G_i identifient les G_i à des quotients séparés de G, on dit quelquefois que les G_i donnent une *approximation* de G.

Considérons un groupe topologique G et une famille (H_i) de sous-groupes distingués fermés, filtrante pour la relation $\supset$. Les morphismes canoniques G → G/H_i définissent un morphisme :

$$f: \quad G \to G' = \varprojlim G/H_i\,;$$

si tout voisinage de e dans G contient l'un des H_i, on démontre que f est un morphisme strict dont le noyau est $\{\bar{e}\}$ et dont l'image est dense dans G'. En particulier, si G est compact, f est un isomorphisme de G sur G' ; un groupe compact totalement discontinu G est isomorphe à la limite projective des groupes finis G/H, où H est un sous-groupe distingué ouvert : on dit que G est un groupe *profini* pour exprimer cette propriété d'approximation par des groupes finis. Par exemple, le groupe de

Galois d'une extension galoisienne infinie d'un corps commutatif est un groupe profini.

Tout groupe localement compact contient un sous-groupe ouvert approchable par des groupes de Lie.

Structures uniformes, groupes complets

Soit G un groupe topologique ; à tout voisinage V de l'élément neutre on associe l'ensemble V_g (resp. l'ensemble V_d) des couples $(x, y) \in G \times G$ tels que $x^{-1}y$ (resp. yx^{-1}) appartienne à V. Lorsque V parcourt le filtre des voisinages de e, les V_g (resp. les V_d) forment le filtre des entourages d'une structure uniforme compatible avec la topologie de G et dite structure uniforme gauche (resp. droite) ; en général les structures uniformes gauche et droite sont distinctes, bien qu'isomorphes par la symétrie $x \mapsto x^{-1}$. Cependant, elles coïncident si G est commutatif ou compact. Si G est séparé et si e a un système fondamental *dénombrable* de voisinages, on peut construire une distance d sur G qui définit sa structure uniforme gauche (resp. droite) et est invariante à gauche (resp. à droite), c'est-à-dire telle que :

$$d(gs, gt) = d(s, t),$$

pour g, $s, t \in G$; ainsi G est *métrisable* (cf. ESPACES MÉTRIQUES).

On dit que G est *complet* si l'une ou l'autre des structures uniformes gauche et droite en fait un espace complet (c'est-à-dire où tout filtre de Cauchy converge). Par exemple, les groupes localement compacts sont toujours complets. Soit $\widehat{G}$ le complété d'un groupe topologique pour sa structure uniforme gauche ; la loi de G se prolonge continûment en une loi de monoïde sur G, mais la symétrie $x \mapsto x^{-1}$, pour $x \in G$, ne se prolonge pas en général, et $\widehat{G}$ n'est pas un groupe. Si G est commutatif, son complété G est un groupe topologique ; il en est de même si la topologie de G est définie par une famille filtrante (H_i) de sous-groupes distingués tels que les quotients G/H_i soient complets. Alors le morphisme :

$$G \rightarrow G' = \varprojlim G/H_i$$

se prolonge en un isomorphisme de G sur G'. Par exemple, la limite projective $\widehat{Z}$ des quotients finis Z/mZ de Z est le complété de Z pour la topologie des sous-groupes d'indice fini (cf. exemple 6) ; c'est le groupe de Galois de la clôture algébrique d'un corps fini, engendré topologiquement par l'élément de Frobenius $a \mapsto a^q$ où q est le cardinal du corps de base. Le complété de Z pour la topologie p-adique (cf. théorie des NOMBRES - Nombre p-adiques) est :

$$Z_p = \varprojlim Z/p^n Z.$$

La théorie des familles sommables et des séries peut se développer dans un groupe topologique complet commutatif noté additivement comme dans R (cf. SÉRIES ET PRODUITS INFINIS).

Voici un résultat technique utile en algèbre commutative. On associe à tout groupe G filtré par des sous-groupes distingués G_n (cf. exemple 6) la famille des groupes quotients (G_n/G_{n+1}), pour $n \in Z$; on obtient ainsi un « groupe gradué », que l'on note gr G. Si $u : G \rightarrow H$ est un homomorphisme de groupes filtrés tel que $u(G_n) \subset H_n$, pour tout n, on désigne par $\mathrm{gr}^n u$ l'homomorphisme de G_n/G_{n+1} dans H_n/H_{n+1} déduit de u par passage au quotient, d'où le morphisme de groupes gradués $\mathrm{gr}\, u : \mathrm{gr}\, G \rightarrow \mathrm{gr}\, H$. On munit G et H des topologies définies par leurs filtrations et on suppose que G est réunion des G_n et H réunion des H_n. Alors :

a) si G est séparé et si gr u est injectif. u est injectif;

b) si G est complet et H séparé, et si gr u est surjectif, u est surjectif (si H est discret, l'hypothèse que G est complet est inutile) ;

c) si G est complet et H séparé, et si gr u est bijectif, u est bijectif.

2. Espaces à groupe d'opérateurs

Soit G un groupe topologique et E un espace topologique. Une loi d'opération de G à gauche sur E est une application *continue* de G X E dans E qui vérifie les axiomes suivants, où $g.x$ désigne l'image de (g, x) dans E :

1. Associativité : quels que soient g, $h \in$ G et $x \in$ E, on a $g.(h.x) = (gh).x$.

2. Quel que soit $x \in$ E, on a $e.x = x$, où e désigne l'élément neutre de G.

On définirait de même les lois de composition a droite.

Pour tout x $\in$ E, l'application *orbitale* $\varphi_x : g \mapsto g.x$ de G dans E, avec g $\in$ G, a pour image l'*orbite* G.x de x, et l'image réciproque de x par φ_x est un sous-groupe G_x de G appelé *le stabilisateur* de x. Deux orbites distinctes dans E sont disjointes ; ainsi les orbites sont les classes pour une relation d'équivalence sur E, et l'on désigne l'espace topologique quotient par E/G (ou par G/E si on veut rappeler que G opère à gauche).

Soit H un sous-groupe d'un groupe topologique G ; il opère à droite sur G par la loi $(s, h) \mapsto sh$, pour $s \in$ G et $h \in$ H. L'orbite sH de s est aussi appeiée sa classe à droite suivant H ; le groupe G opère à gauche sur le quotient G/H par la loi $(g, s\text{H}) \mapsto gs\text{H}$, pour g, s $\in$ G, et cela de façon transitive (il y a une seule orbite). Lorsque G est métrisable complet, on démontre que G/H est encore métrisable complet pour tout sous-groupe fermé H de G.

On dit que G opère *proprement* sur E si l'application (s, x) $\mapsto$ (x, $s.x$) de G X E dans E X E est propre. Alors, les applications orbitales sont propres, les orbites sont *fermées* et les stabilisateurs sont *compacts* ; enfin, pour tout $x \in$ E, l'application orbitale définit un *isomorphisme* de G/G_x sur G.x (en particulier si G opère proprement et transitivement dans E, l'espace E est isomorphe à G/G_x pour tout $x \in$ E). Lorsque G est localement compact et E séparé, pour que G opère proprement il faut et il suffit que la condition suivante soit satisfaite : Quels que soient les points x et y de E, il existe des voisinages V et W de x et de y respectivement tels que l'ensemble des $s \in$ G pour lesquels on a $s.\text{V} \cap \text{W} \neq \emptyset$ soit un ensemble *relativement compact,* donc fini dans le cas particulier où G est discret. Dans ce cas, si E est localement compact, on a encore le critère suivant pour une loi d'opération propre : Pour toute partie compacte K de E, l'ensemble des $s \in$ G pour lesquels on a $s.\text{K} \cap \text{K} \neq \emptyset$ est un ensemble fini.

3. Anneaux topologiques

On appelle anneau topologique un groupe topologique A commutatif, noté additivement, muni d'une loi multiplicative *continue* A X A $\longrightarrow$ A qui en fait un anneau (cf. ANNEAUX ET ALGÈBRES). Un morphisme d'anneaux topologiques est une application continue qui est un homomorphisme d'anneaux.

Exemples

1. Les anneaux discrets sont des anneaux topologiques.

2. Les corps R et C des nombres réels et complexes, les anneaux $M_m(R)$ et $M_m(C)$ des matrices carrées d'ordre *m* à coefficients réels ou complexes, les algèbres normées, réelles ou complexes, sont des anneaux topologiques.

3. Les corps p-adiques $\mathbf{Q}_p$ et l'anneau des adèles A (cf. théorie des NOMBRES - Nombres algébriques) sont des anneaux topologiques.

4. On peut faire d'un anneau A un anneau topologique, un « anneau linéairement topologisé », au moyen de la topologie définie par une famille filtrante décroissante d'idéaux bilatères (cf. chap. 1, exemple 6).

On peut encore associer une topologie compatible avec la structure d'anneau à toute *filtration* (A_n), avec $n \in Z$, suite décroissante de sous-groupes du groupe additif de A telle que $A_m A_n \subset A_{m+n}$ pour $m, n \in Z$ et $1 \in A_0$. Par exemple, si m est un idéal bilatère de A, il définit une filtration dite m-adique, pour laquelle $A_n = \mathfrak{m}^n$ si $n \geqslant 0$ et $A_n = A$ si $n < 0$; on en déduit la *topologie* m-adique sur A. Pour A = Z et m= $p\mathbf{Z}$, où p est nombre premier, on retrouve la topologie p-adique ; comme autre exemple, signalons le cas où $A = K[X_1, X_2, ..., X_r]$ est un anneau de polynômes à coefficients dans un anneau commutatif K et où m = $(X_1, X_2, \dots X_r)$.

Soit A un anneau topologique ; sa loi multiplicative se prolonge continûment en $\hat{A} \times \hat{A} \rightarrow \hat{A}$ et le complété $\hat{A}$ est encore un anneau topologique. Par exemple. le complété $\mathbf{Z}_p$ de Z pour la topologie p-adique est un anneau topologique ; on peut en dire autant du complété :

$$K[[X_1, X_2, \dots, X_r]] = \varprojlim_n K[X_1, X_2, \dots X_r]/\mathfrak{m}^n$$

de $K[X_1, X_2, \dots X_r]$ pour la topologie m-adique, où m = $(X_1, X_2, \dots X_r)$. Cet anneau filtré complet n'est autre que l'anneau des *séries formelles* en les X, à coefficients dans K. Soit A un anneau commutatif et m un idéal *maximal* de A ; le complété de A pour la topologie $\mathfrak{m}$-adique est un anneau *local* d'idéal maximal $\hat{\mathfrak{m}}$ (c'est le cas de $\mathbf{Z}_p$ et de l'anneau des séries formelles si K est un corps).

Soit A un anneau commutatif complet pour une topologie définie par une famille filtrante d'idéaux a ; le complété de l'anneau de polynômes $A[X_1, X_2, \dots X_r]$ pour la topologie définie par les idéaux $\mathfrak{a}[X_1, X_2, \dots X_r]$ s'identifie au sous-anneau de $A[[X_1, X_2, \dots X_r]]$ formé des séries formelles dont les coefficients tendent vers 0 dans A ; cet anneau est appelé l'anneau des séries formelles restreintes. Le *lemme de Hensel* (cf. théorie des NOMBRES - Nombres padiques, chap. 2) a une formulation générale dans ce cadre.

Soit A un anneau commutatif et m un idéal de A ; pour tout A-module M, la filtration m-adique $(\mathfrak{m}^n M)$, avec $n \in N$, fait de M un groupe topologique additif tel que la loi externe $A \times M \rightarrow M$ soit continue (« module topologique »). Le *théorème de Krull* affirme que, dans le cas où A est noethérien et M de type fini, la topologie induite par celle de M sur un sous-module quelconque N est identique à la topologie m-adique de N, définie par la filtration $(\mathfrak{m}^n N)$. Si m est contenu dans *le radical* de A (intersection des idéaux maximaux), tout A-module de type fini est *séparé* pour la topologie m-adique et le complété $\widehat{M}$ d'un tel module M s'identifie à $\hat{A} \otimes_A M$; les suites exactes de modules de type fini le restent après complétion ; autrement dit, $\hat{A}$ est *plat* sur A.

4. Corps topologiques

On appelle corps topologique un anneau topologique K qui est un corps et dans lequel l'application $x \mapsto x^{1}$, pour x $\in$ K et x $\neq$ 0, est continue. Les plus importants des corps topologiques sont les *corps valués*, dont la structure est définie par une *valeur absolue,* c'est-à-dire une application x $\mapsto$ x de K dans $\mathbf{R}_+$ vérifiant :

$$(\mathrm{VA}_1) \qquad |x| = 0 \Leftrightarrow x = 0,$$
$$(\mathrm{VA}_2) \qquad |xy| = |x| . |y|,$$
$$(\mathrm{VA}_3) \qquad |x + y| \leqslant |x| + |y|,$$

quels que soient x, y $\in$ K. Une valeur absolue définit une distance $d(x, y)$ = y x |, donc une topologie. On dit que la valeur absolue est *ultramétrique* si elle vérifie la condition suivante, plus forte que (VA_3) :

$$|x + y| \leqslant \sup(|x|, |y|),$$

quels que soient x, y $\in$ K ; dans ce cas, l'application v : $x \mapsto \log 1/$ x de K* = K {0} dans R est une *valuation*, c'est-à-dire vérifie, pour x, $y \in$ K*, les deux conditions :

$$(\mathrm{VL}_1) \qquad v(xy) = v(x) + v(y),$$
$$(\mathrm{VL}_2) \qquad v(x + y) \geqslant \inf(v(x), v(y)).$$

Inversement. toute valuation définit une valeur absolue par la formule $x = a^{v(x)}$, où a est un nombre réel choisi dans l'intervalle]0, 1[.

Exemples

1. Sur le corps Q des nombres rationnels, les valeurs absolues sont de l'un des trois types suivants :

u) la valeur absolue impropre, qui vaut 0 en 0 et 1 ailleurs ;

b) les valeurs absolues p-adiques :

$$|x|_p = a^{v_p(x)},$$

où l'on a 0 < a < 1 et où v_p est la valuation p-adique (cf. théorie des NOMBRES - Nombres p-adiques) ;

c) les applications $x \mapsto x^{s}$, où x est la valeur absolue usuelle et où l'on a $0 < s \leqslant 1$. Ces valeurs absolues définissent la même topologie que la valeur absolue usuelle.

2. Les corps topologiques localement compacts sont des corps valués (cf. théorie des NOMBRES - Nombres algébriques).

3. Si K est un corps valué non ultramétrique, il existe un isomorphismej de K sur un sous-corps partout dense de l'un des corps R, C ou le corps H des quaternions et un nombre réel $s \in]0, 1]$ tels que x = $j(x)|^{s}$, pour $x \in$ K (théorème d'Ostrowski).

Le complété d'un corps topologique K est un anneau qui n'est pas un corps en général ; cependant c'est un corps dans le cas où K est valué.

CHRISTIAN HOUZEL

N. BOURBAKI, *Éléments de mathématiques*, livre III : *Topologie générale*, Hermann, Paris, 1960, nouv. éd. 1968 ; livre VIII : *Algèbre commutative*, *ibid.*, 1961-1964, nouv. éd. 1965-1969 / D. MONTGOMERY & L. ZIPPIN, *Topological Transformation Groups*, Interscience Publishers, New York, 1955 / L. S. PONTRIAGIN, *Topological Groups*, Princeton Univ. Press, Princeton, 1946.

TOPOLOGIQUES ESPACES VECTORIELS

La théorie des espaces normés, développée par S. Banach et ses élèves, s'est vite révélée insuffisante pour les besoins de

l'analyse fonctionnelle où interviennent de nombreux espaces vectoriels munis d'une topologie qui n'est pas déduite d'une norme. Les espaces vectoriels topologiques et leurs variantes définis dans cet article sont des généralisations des espaces normés ; la convexité y joue un rôle essentiel.



1. Définition et exemples

Soit K un corps valué complet non discret, par exemple R ou C (cf. algèbre TOPOLOGIQUE). Une topologie sur un K-espace vectoriel E est dite *vectorielle* si les applications $(x, y) \mapsto x + y$ et $(\lambda, x) \mapsto \lambda x$ de E X E dans E et de K X E dans E sont continues ; on appelle espace vectoriel topologique sur K un K-espace vectoriel muni d'une topologie vectorielle. Un tel espace E est, pour sa loi additive, un groupe commutatif topologique, et sa topologie est donc déterminée par le filtre $\mathscr{V}$ des voisinages de l'élément neutre 0. Rappelons que ce filtre vérifie les conditions suivantes (écrites ici en notation additive) :

(GV_1) Quel que soit $U \in \mathscr{V}$, il existe $V \in \mathscr{V}$ tel que $V + V \subset U$;

(GV_2) Quel que soit $U \in \mathscr{V}$, on a $-U \in \mathscr{V}$.

En exprimant la continuité de l'application $(\lambda, x) \mapsto \lambda x$, on trouve les conditions suivantes, dont la première renforce (GV_2) :

(EV_1) Quels que soient $U \in \mathscr{V}$ et $\lambda \in K$, on a $\lambda U \in \mathscr{V}$; de plus, on a :

$$\bigcup_{\lambda \in K} \lambda U = E.$$

(EV_2) Quel que soit $U \in \mathscr{V}$, il existe $V \in \mathscr{V}$, contenu dans U et *équilibré,* c'est-à-dire tel que :

$$\bigcup_{|\lambda| \leqslant 1} \lambda V \subset V.$$

Inversement, considérons un K-espace vectoriel E et un filtre $\mathscr{V}$ sur E possédant les propriétés (GV_1), (EV_2) et (EV_1) ; il existe sur E une topologie vectorielle et une seule pour laquelle $\mathscr{V}$ est le filtre des voisinages de 0.

Un espace vectoriel topologique E a une structure uniforme naturelle dont les entourages sont les ensembles :

$$\tilde{V} = \{(x, y) \in E \times E \,|\, y - x \in V\}, \qquad V \in \mathscr{V}.$$

Pour que E soit séparé, il faut et il suffit que $\{0\}$ soit fermé ; si, de plus, le filtre $\mathscr{V}$ admet une base dénombrable, il existe une distance invariante par translation qui définit la structure uniforme de E. Les lois interne et externe de E se prolongent par continuité au complété $\hat{E}$ et en font un espace vectoriel topologique sur K.

Les espaces vectoriels topologiques que l'on rencontre en analyse sont le plus souvent *localement convexes,* c'est-à-dire que leurs filtres $\mathscr{V}$ de voisinages de 0 vérifient la condition suivante, plus forte que (EV_1) :

(EV'_1) Quel que soit $U \in \mathscr{V}$, il existe $V \in \mathscr{V}$ qui est *disqué* et contenu dans U.

Une partie V d'un espace vectoriel E est dite *disquée* si, pour toute famille finie (λ_i), avec $i \in I$, de scalaires, on a :

$$\sum_{i \in I} |\lambda_i| \leqslant 1 \Rightarrow \sum_{i \in I} \lambda_i V \subset V;$$

on dit encore que V est un *disque.* Si K = R (resp. C), les disques de E sont les parties *convexes* (cf. CONVEXITÉ) et symétriques (resp. équilibrées), d'où le terme « localement convexe ». Notons que les

conditions (EV,) et (EV,) impliquent (GV,) car, si V est disqué, on a $V \supset \lambda V + AV$, avec $A \in K$, choisi tel que $0 < \lambda \leqslant 1/2$ (rappelons que K n'est pas discret). Le complété d'un espace vectoriel topologique localement convexe (en abrégé e.l.c.) est encore localement convexe ; on appelle *espace de Fréchet* un e.l.c. métrisable et complet.

Exemples

1. Les espaces « numériques » K^m, avec la topologie produit, sont des e.l.c. complets ; on démontre (par récurrence sur m) que la seule topologie vectorielle *séparée* sur K^m est la topologie produit.

2. Tout espace vectoriel normé E est un e.l.c. ; un système fondamental de voisinages de 0 dans E est formé des homothétiques de la houle *unité :*

$$B = \{x \in E \mid \|x\| \leqslant 11.$$

Comme cas particuliers, on trouve l'espace e(X) des fonctions numériques continues sur un espace topologique compact X (avec la topologie de la convergence uniforme dans X), l'espace l^p des suites de scalaires de puissance p-ième sommabie et l'espace $L^p(X, \mu)$ des classes de fonctions de puissance p-ième intégrable dans un espace mesuré (X, μ), avec $p \geqslant 1$ et K = C ; ces espaces sont complets, c'est-a-dire que ce sont des espaces de Banach. Lorsque $0 < p < 1$, on peut encore définir un espace vectoriel topologique L"(X, µ), qui est métrisable et complet, mais non localement convexe.

3. Si X est un espace topologique séparé, on fait un e.l.c. de l'espace e(X) des fonctions numériques continues dans X en le munissant de la topologie de la convergence uniforme sur tout compact de X ; un système fondamental de voisinages de 0 pour cette topologie est formé des ensembles :

$$V_{C,\varepsilon} = \{f \in \mathcal{C}(X) \mid \|f\|_C = \sup_{x \in C} |f(x)| \leqslant \varepsilon\},$$

où C est un compact quelconque de X et où $\varepsilon \in$ RT. Cet espace est complet si X est localement compact ou métrisable ; il est métrisable s'il existe une suite (C,) de compacts de X telle que tout compact de X soit contenu dans l'un des C,,.

En utilisant les ensembles finis F C X au lieu des compacts C, on définit d'une manière analogue la topologie de la convergence simple qui est aussi vectorielle et localement convexe, mais non complète sur e(X) si X est infini.

4. Soit X une variété différentiable, de classe $m \leqslant +\infty$, par exemple un ouvert de $\mathbf{R}^n$. Sur l'espace $\mathcal{E}^n(X)$ des fonctions numériques de classe m dans X, on définit une topologie vectorielle localement convexe en prenant comme système fondamental de voisinages de 0 la famille des ensembles :

$$V_{\Delta,C,\varepsilon} = \{f \in \mathcal{E}n(X) \mid \forall D \in \Delta, \|Df\|_C \leqslant \varepsilon\},$$

où A est un ensemble fini d'opérateurs de dérivation d'ordre $\leqslant m$, où C est un compact de X et où l'on a $\varepsilon > 0$. L'e.l.c. $\mathcal{E}^n(X)$ ainsi obtenu est complet ; il est métrisable si X est dénombrable à l'infini (ce qui est le cas pour un ouvert de R") et normable si X est compacte et m fini. On écrit $\mathcal{E}(X)$ pour C"(X).

5. La construction de l'exemple 4 admet plusieurs variantes ; par exemple, si X est un ouvert de $\mathbf{R}^n$ et sip $\geqslant 1$, on définit un espace de Fréchet $\mathcal{D}_{L^p}(X)$ dont les éléments sont les fonctions indéfiniment dérivables dans X dont toutes les dérivées sont de puissancep-ième intégrable, avec la topologie définie par le système fondamental de voisinages de 0 :

$$V_{\Delta,\varepsilon} = \{f \in \mathcal{D}_{L^p}(X) \mid \vee \text{ DE A}, \|Df\|_p \leqslant \varepsilon\},$$

où A est un ensemble fini d'opérateurs de dérivation et où l'on a $\varepsilon > 0$, la norme dans $L^p(X)$ étant désignée par $\|\cdot\|_p$. Citons encore l'espace S des fonctions indéfiniment dérivables à décroissance rapide dans $\mathbf{R}^n$, qui est aussi un espace de Fréchet ; ses éléments sont les fonctions indéfiniment dérivables dont toutes les dérivées décroissent à l'infini de $\mathbf{R}^n$ plus vite que toute puissance de $1/\|x\|$, une norme étant choisie dans $\mathbf{R}^n$; un système fondamental de voisinages de 0 est donné par :

$$V_{\Delta,k,\varepsilon} = \{f \in \mathcal{S} \mid \forall\, \mathrm{D} \in \Delta,\ \sup_{x \in \mathbf{R}^n} (1 + \|x\|^2)^k |\mathrm{D}f(x)| \leqslant \varepsilon\},$$

où A est un ensemble fini d'opérateurs de dérivation, k un entier naturel et $\varepsilon > 0$.

6. Soit U un ouvert de $\mathbf{C}^n$ et $\mathcal{O}(U)$ l'espace vectoriel des fonctions holomorphes dans U. La topologie de la convergence uniforme sur tout compact de U (cf. exemple 3) en fait un espace de Fréchet.

7. Soit $S(X, \mu)$ l'espace vectoriel des classes de fonctions p-mesurables sur un espace mesuré (X, μ) ; on le munit d'une topologie vectorielle en prenant comme système fondamental de voisinages de 0 la famille des ensembles $V_{A,\delta,\varepsilon}$:

$$\{f \in S(X,\mu) \mid \mu^*(A \cap f^{-1}(\complement[-\varepsilon, \varepsilon])) \leqslant \delta\},$$

où A est une partie μ-intégrable de X et où δ et ε sont des nombres réels > 0. La topologie ainsi définie est appelée celle de la *convergence en mesure* et elle est importante en calcul des probabilités ; elle n'est pas localement convexe, mais elle fait de $S(X, \mu)$ un espace vectoriel topologique complet.

2. Limites projectives et limites inductives

Soit (E_i), pour $i \in I$, une famille d'espaces vectoriels topologiques, E un espace vectoriel et, pour chaque i, une application linéaire $f_i : E \to E_i$; la moins fine des topologies sur E pour lesquelles les f_i sont continues est vectorielle : on l'appelle la topologie *initiale* pour les f_i. Si les E_i sont tous localement convexes, il en est de même de E muni de la topologie initiale.

Lorsque $I = \{0\}$ est réduit à un élément et que $f_0 : E \to E_0$ est l'injection d'un sous-espace vectoriel, la topologie initiale n'est autre que la topologie induite ; elle est séparée (resp. métrisable) si celle de E_0 l'est. Si E_0 est complet et E fermé, E est aussi complet ; l'adhérence d'un sous-espace vectoriel est encore un sous-espace vectoriel.

Sur le produit $E = \prod E_i$ d'une famille (E_i), pour $i \in I$, d'espaces vectoriels topologiques, la topologie initiale pour les projections $\mathrm{pr}_i : E \to E_i$ est la topologie produit. Considérons plus généralement un système projectif (E_i, f_{ij}) d'espaces vectoriels topologiques, avec des morphismes de transition $f_{ij} : E_j \to E_i$ linéaires continus ; la *limite projective* $E = \varprojlim E_i$, c'est-à-dire le sous-espace de $\prod E_i$ formé des (x_i) tels que $f_{ij}(x_j) = x_i$ pour tout morphisme de transition f_{ij}, est munie de la topologie initiale pour les applications $f_i : E \to E_i$ induites par les projections. Lorsque les E_i sont séparés (resp. complets), il en est de même de E ; si les E_i sont métrisables et si I est dénombrable, E est métrisable.

À tout disque V d'un espace vectoriel E on associe la semi-norme p_V *jauge* de V (cf. CONVEXITÉ), telle que :

$$p_V(x) = \inf\{|\lambda| \mid x \in \lambda V\}, \qquad x \in E,$$

et l'espace semi-normé E_V, qui est le sous-espace engendré par V muni de la semi-norme p_V. Supposons E muni d'une topologie localement convexe et prenons pour V un voisinage disqué de 0 ; alors, en

vertu de la condition (EV,), l'espace Ev est E tout entier (on dit que V est *absorbant*) et, lorsque V varie, la topologie de E est initiale pour les applications E ⟶ Ev égales à l'application identique. On dit que cette topologie est définie par la famille de semi-normes $(p_V)_V$; les espaces E et lim E_V sont alors isomorphes en tant qu'espaces topologiques. Lorsque E est séparé, on en déduit qu'il est limite projective des espaces normés E_V, en désignant par E_V l'espace séparé associé à Ev ; lorsque E est complet et que K est R ou C ou un corps « maximalement complet », on voit à l'aide du théorème de Hahn-Banach (cf. CONVEXITÉ) que E est limite projective des espaces de Banach $\hat{E}_V$.

De façon duale, on peut considérer la *topologie finale* sur un espace vectoriel E pour une famille d'applications linéaires f_i : $E_i \rightarrow$ E définies dans des espaces vectoriels topologiques E_i : c'est la plus fine des *topologies vectorielles* sur E rendant continues les f_i ; elle n'est pas en général localement convexe, même si tous les E_i le sont, et, dans ce cas, on considère plutôt la topologie localement convexe la plus fine rendant les f_i continues, qui est moins fine que la précédente. En particulier, on définit ainsi l'espace vectoriel topologique (resp. l'e.l.c.) limite inductive d'un système inductif (E_i, f_{ij}) d'espaces vectoriels topologiques (resp. d'e.l.c.).

Soit E, un espace vectoriel topologique et E_0 un sous-espace de E, ; sur le quotient E = E_1/E_0, la topologie vectorielle finale pour l'application canonique E, ⟶ E est la topologie quotient et les voisinages de 0 dans E sont les images des voisinages de 0 dans E,. Si E, est localement convexe, il en est de mëme de E ; pour que E soit séparé, il faut et il suffit que E_0 soit fermé dans E,. Si E_1 est métrisable et complet et si E_0 est fermé, alors E est métrisable et complet (cf. algèbre TOPOLOGIQUE). Toute application linéaire continue f : E ⟶ F d'un espace vectoriel topologique dans un autre admet une décomposition canonique :

$$E \rightarrow E/\mathrm{Ker} f \xrightarrow{\tilde{f}} f(E) \hookrightarrow F\,;$$

l'application linéaire $\tilde{f}$ est bijective mais n'est pas un homéomorphisme en général : on dit que f est *stricte* si $\tilde{f}$ est un isomorphisme d'espaces vectoriels topologiques. Le *théorème d'homomorphisme de Banach* s'énonce de la manière suivante.

Théorème. Soit E et F des espaces vectoriels topologiques *métrisables* et ***complets*** et fune application linéaire continue de E dans F. Si $f(E)$ n'est pas ***maigre***, par exemple si f est surjective (théorème de Baire ; cf. espaces vectoriels NORMÉS, chap. 4), alors f est un *épimorphisme strict*, identifiant F à un quotient de E.

Corollaire (Théorème du graphe ***fermé).*** Soit g : F- E une application linéaire d'un espace vectoriel topologique métrisable complet dans un autre. Pour que g soit continue, il faut et il suffit que, pour toute suite (y_n) convergeant vers 0 dans F et telle que $(g(y_n))$ converge dans E, on ait lim $g(y_n) = 0$.

En effet, la condition signifie que le graphe Γ_g de g est un sous-espace fermé de F X E, donc un espace métrisable et complet; la projection $\Gamma_g \rightarrow F$, qui est bijective, est donc un isomorphisme et g s'obtient en composant l'isomorphisme réciproque avec l'autre projection : $\Gamma_g \rightarrow E$.

Il existe diverses généralisations de ces résultats, où F est remplacé par un espace localement convexe séparé, limite inductive d'espaces de Banach, et E par un espace localement convexe souslinien (L.

Schwartz-A. Martineau) ou par un espace admettant un « réseau absorbant » (M. De Wilde).

3. Bornologies

Pour comprendre de façon claire et naturelle la théorie des espaces vectoriels topologiques, il faut étudier simultanément une structure voisine, celle d'espace vectoriel bornologique, que les traités classiques laissent malheureusement de côté au prix d'un certain nombre d'obscurités et de maladresses.

Une *bornologie* sur un ensemble E est un ensemble $\mathscr{B}$ de parties de E vérifiant les axiomes suivants :

(B,) Quels que soient A $\in \mathscr{B}$ et A' C A, on a A' $\in \mathscr{B}$;

(B_2) Quels que soient A,, $A_2 \in \mathscr{B}$, on a A, $\cup A_2 \in \mathscr{B}$;

(B_3) La réunion des A, pour A $\in \mathscr{B}$, est égale à E.

Un ensemble E muni d'une bornologie s'appelle un *espace hornologique* et les éléments de la ~~bornologie de E s'appellent~~ les *parties bornées* de E ; une application f: E $\rightarrow$ F d'un espace bornologique dans un autre est dite *bornée* si elle transforme les parties bornées de E en parties bornées de F. Le corps K est, comme tout espace métrique, muni d'une bornologie naturelle dont les éléments sont les parties de diamètre fini. Si (E,), pour $i \in$ I, est une famille d'espaces bornologiques, la bornologie produit sur E = PE, a pour éléments les parties dont toutes les projections sont bornées ; par exemple, K X K est un espace bornologique pour la bornologie produit et l'addition et la multiplication de K sont des applications bornées de K X K dans K. Une bornologie $\mathscr{B}$ sur un K-espace vectoriel E est dite *vectorielle* si les applications $(x, y) \mapsto x + y$ et $(\lambda, x) \mapsto \lambda x$ de E X E dans E et de K $\times$ E dans E sont bornées. Cela équivaut aux conditions suivantes :

(EB_1) Quels que soient A,, $A_2 \in \mathscr{B}$, on a $A_1 + A_2 \in \mathscr{B}$;

(EB_2) Quels que soient A $\in \mathscr{B}$ et $\lambda \in$ K, on a $\lambda A \in \mathscr{B}$;

(EB_3) Quel que soit A $\in \mathscr{B}$, l'« enveloppe équilibrée de A », c'est-à-dire la réunion des ensembles AA, pour A $\leqslant$ 1, appartient à $\mathscr{B}$.

Un espace vectoriel muni d'une bornologie vectorielle s'appelle un espace vectoriel bornologique. Les plus fréquents des espaces vectoriels bornologiques sont de *type* convexe, vérifiant la condition plus forte que (EB,) :

(EB'_3) L'enveloppe disquée d'un borné est bornée.

Remarquons que (B_2), (EB_2) et (EB',) impliquent (EB,).

On dit qu'un espace vectoriel bornologique E est *séparé* si le seul sous-espace vectoriel borné de E est $\{0\}$.

Exemples

1. La bornologie produit sur K^m est la seule bornologie vectorielle séparée ; elle est de type convexe.

2. Tout espace vectoriel normé E a une bornologie de type convexe dont les bornés sont les parties de diamètre fini, c'est-à-dire contenues dans une boule.

3. Sur un espace vectoriel quelconque E, la bornologie vectorielle *la plus fine* est formée des parties contenues dans un sous-espace de dimension finie et bornées dans ce sous-espace ; elle est de type convexe et séparée.

4. Soit X une variété différentiable de classe $m \leqslant +\infty$; sur l'espace 'd)"(X) des fonctions numériques m fois différentiables à support compact dans X, on définit

une bornologie vectorielle de type convexe, dont les éléments sont les ensembles de fonctions uniformément bornées, ainsi que toutes leurs dérivées d'ordre $\leqslant m$ et nulles en dehors d'une même compact de X.

Voici un autre exemple d'espace vectoriel bornologique de type convexe (en abrégé e.b.c.) de fonctions différentiables : l'espace des fonctions indéfiniment différentiables *à croissance lente* dans $\mathbf{R}^n$. Ses éléments sont les fonctions f indéfiniment différentiables dont toutes les dérivées ont au plus une croissance polynomiale à l'infini de $\mathbf{R}^n$; les bornés sont les parties contenues dans l'un des ensembles :

$$B_{(P_D)} = \{f ; \forall D, \forall x, |Df(x)| \leqslant |P_D(x)|\},$$

où (P_D) est une famille de polynômes indexée par l'ensemble des opérateurs de dérivation D ; on dit que les $B_{(P_D)}$ forment un système fondamental de bornés.

5. Soit U un ouvert de K^n et $\mathcal{O}(U)$ l'espace des fonctions analytiques dans U, c'est-à-dire localement développables en série entière convergente dans U. Si V est un polydisque de centre a et de rayon r contenu dans U et si f est une fonction analytique dans U dont le développement de Taylor en a est :

$$\Sigma c_\nu (z - a)^\nu,$$

on pose :

$$\|f\|_V = \Sigma |c_\nu| r^\nu ;$$

on définit alors sur $\mathcal{O}(U)$ une bornologie de type convexe en prenant comme système fondamental de bornés les ensembles :

$$B_{\mathcal{U}, \mathcal{M}} = \{f \in \mathcal{O}(U) \mid \forall i, \ \|f\|_{U_i} \leqslant M_i\},$$

où $\mathcal{U} = (U_i)$, pour $i \in I$, est un recouvrement de U des polydisques et où $\mathcal{M} = (M_i)$, pour $i \in I$, est une famille de constantes.

On définit sans peine les notions de bornologie initiale et de bornologie vectorielle finale pour une famille d'applications linéaires, comme au chapitre 2. On peut ainsi parler de sous-espaces et de limites projectives bornologiques, ainsi que d'espaces quotients et de limites inductives bornologiques. Considérons un e.b.c. E ; sa bornologie est finale pour les injections $E_A \rightarrow E$ où A décrit l'ensemble des disques bornés de E et où E_A est le sous-espace engendré par A muni de la semi-norme p_A, jauge de A. On a un isomorphisme :

$$\varinjlim_A E_A \rightleftarrows E$$

d'espaces vectoriels bornologiques ; pour que E soit séparé, il faut et il suffit que chaque E_A le soit, c'est-à-dire soit un espace normé. Par exemple, la bornologie la plus fine sur un espace vectoriel E est limite inductive des bornologies séparées sur les sous-espaces de dimension finie de E.

On dit qu'un e.b.c. E est *complet* s'il existe un système fondamental de disques bornés A tels que E_A soit complet ; cela revient à dire que E est séparé et limite inductive d'espaces de Banach. Tout e.b.c. E admet un complété $\hat{E}$ qui est le « séparé associé » à :

$$\varinjlim_A \hat{E}_A ;$$

l'application canonique de E dans $\hat{E}$ n'est pas injective en général, même si E est séparé : il existe des e.b.c. séparés non nuls dont le complété est nul, comme l'a montré L. Waelbroeck. Toute limite projective et toute limite inductive séparée d'e.b.c. complets sont complètes. Les exemples donnés sont des e.b.c. complets.

On dit qu'une suite (x_n) de points d'un espace vectoriel bornologique E converge vers une limite x *au* sens *de Mackey* s'il existe une suite bornée (y_n) dans E et une suite (λ_n) tendant vers 0 dans K telles que $x_n \quad x = \lambda_n y_n$. Une partie X de E est dite ***fermée*** (au sens de Mackey) si toute limite (au sens de Mackey) de points de X appartient à X ; pour qu'un sous-espace F de E soit fermé, il faut et il suffit que le quotient E/F soit séparé. Pour que E soit séparé, il faut et il suffit que toute suite convergente dans E ait une seule limite, ou encore que la diagonale Δ_E soit fermée dans E X E. Lorsque E est de type convexe, la convergence au sens de Mackey dans E est équivalente à la convergence dans l'un des espaces semi-normés E_A ou A est un disque borné dans E. Dans un e.b.c. complet E, toute suite de « Cauchy-Mackey » (x,), telle que $(x_m - x_n)$ tende vers 0 au sens de Mackey pour ***m,*** $n \rightarrow \infty$, est convergente ; mais cette condition n'est pas suffisante pour assurer que E est complet.

4. Espaces d'applications linéaires et produits tensoriels

Soit E et F des espaces vectoriels topologiques ; on désigne par C(E, F) l'espace vectoriel des applications linéaires continues de E dans F, muni de la bornologie vectorielle dont les éléments sont les ensembles *équicontinus* d'applications linéaires : un ensemble H d'applications linéaires est équicontinu s'il l'est en 0, c'est-à-dire si, pour tout voisinage V de 0 dans F, il existe un voisinage U de 0 dans E tel que :

$$H(U) = \bigcup_{f \in H} f(U) \subset V.$$

Lorsque E = K, l'application $f \mapsto f(1)$ identifie $\mathcal{L}(K, F)$ à F et permet de munir F d'une bornologie vectorielle, la bornologie canonique ou bornologie de von Neumann ; ses éléments sont les parties A de F qui sont *absorbées* par chaque voisinage V de 0 dans F (il existe $\lambda \in$ K tel que A C AV), et on note ^bF l'espace F muni de cette bornologie. L'adhérence dans F d'un tel borné est encore bornée ; pour que A C F soit borné, il faut et il suffit que toute partie dénombrable A soit bornée. Si U est un ouvert de $\mathbb{C}^n$, la bornologie canonique de l'espace de Fréchet $\mathcal{O}(U)$ des fonctions holomorphes dans U (chap. 1, exemple 6) n'est autre que celle qu'on a définie dans l'exemple 5 du chapitre 3.

Soit maintenant E et F des espaces vectoriels bornologiques ; on fait de l'espace $\mathcal{L}(E, F)$ des applications linéaires bornées de E dans F un espace vectoriel bornologique en prenant comme bornés les ensembles H *équibornés* d'applications linéaires, tels que, pour tout borné A de E. il existe un borné B de F, avec H(A) C B. On associe encore un espace vectoriel bornologique C(E, F) d'applications linéaires au couple d'un espace vectoriel topologique E et d'un espace vectoriel bornologique F ; ses éléments sont les applications linéaires bornantes (c'est-à-dire transformant un certain voisinage de 0 dans E en un borné de F) et ses bornés sont les ensembles H *équibornants* (pour lesquels il existe un voisinage U de 0 dans E et un borné B de F tels que H(U) C B).

Considérons enfin un espace vectoriel bornologique E et un espace vectoriel topologique F. Sur Horn, (E, F), la topologie de la convergence uniforme dans les bornés de E a comme voisinages de 0 les ensembles :

$$T(A, V) = \{f | f(A) \subset V\},$$

où A est un borné de E et V un voisinage de 0 dans F ; ces voisinages engendrent le sous-espace L(E, F) des applications linéaires bornées de E dans ${}^bF$. Sur l'espace $\mathcal{L}(E, F)$ ainsi défini, on voit que la topologie précédente est vectorielle.

À tout couple (E, F) d'espaces vectoriels munis de structures topologiques ou bornologiques nous avons ainsi associé un espace d'applications linéaires C(E, F) qui est bornologique, sauf dans le cas où E est bornologique et F topologique ; dans ce dernier cas, $\mathcal{L}(E, F)$ est topologique. Lorsque F est de type convexe (resp. séparé, resp. complet), il en est de même de C(E, F). En particulier, si F est un e.l.c., ${}^bF$ est de type convexe ; pour que ${}^bF$ soit séparé, il faut et il suffit que F le soit; si F est complet, il en est de même de ${}^bF$, mais la réciproque n'est pas vraie en général (cependant elle est vraie si F est supposé *métrisable*). L'espace C(E, F) dépend fonctoriellement de E et de F ; si (E_i) est un système inductif et (F_j) un système projectif, le morphisme canonique :

$$\mathcal{L}(\varinjlim E_i, \varprojlim F_j) \to \varprojlim \mathcal{L}(E_i, F_j)$$

est un isomorphisme, à condition que le système (E_i) soit fini s'il s'agit d'espaces vectoriels topologiques et que les deux systèmes soient finis si les E_i sont topologiques et les F_j bornologiques. Par exemple, le foncteur $F \to {}^bF$ commute aux limites projectives ; il commute aussi aux limites inductives *strictes*, c'est-à-dire dénombrables et dont les morphismes de transition sont des monomorphismes stricts d'images fermées.

Considérons maintenant trois espaces E, F et G, avec des structures topologiques ou bornologiques. On leur associe d'une manière analogue un espace $\mathcal{B}(E, F ; G)$ d'applications bilinéaires de E X F dans G, avec une structure bornologique en général. Lorsque E, F et G sont topologiques (resp. bornologiques), on prend les applications bilinéaires continues (resp. bornées), avec comme bornés les ensembles équicontinus (resp. équibornés). Dans le cas bornologique (mais pas dans le cas topologique !), on a des isomorphismes canoniques :

$$\mathcal{L}(E, \mathcal{L}(F, G)) \simeq \mathcal{B}(E, F ; G) \simeq \mathcal{L}(F, \mathcal{L}(E, G))$$

Considérons enfin des espaces vectoriels topologiques E et G et un espace vectoriel bornologique F ; on vérifie facilement que la bijection canonique :

$$\mathrm{Hom}_K(E, \mathrm{Hom}_K(F, G)) \rightleftarrows \mathrm{Hom}_K(F, \mathrm{Hom}_K(E, G))$$

induit un isomorphisme d'espaces vectoriels bornologiques :

$$\mathcal{L}(E, \mathcal{L}(F, G)) \rightleftarrows \mathcal{L}(F, \mathcal{L}(E, G)).$$

Les applications bilinéaires φ : $E \times F \to G$ correspondant aux éléments de ces espaces sont caractérisées par la propriété suivante : Pour tout borné B de F et tout voisinage V de 0 dans G, il existe un voisinage U de 0 dans E tel que $\varphi(U \times B) \subset V$; on dit que ces applications bilinéaires sont *hypocontinues.* Par exemple, si E_1 est un espace vectoriel bornologique et si E_2 et E_3 sont des espaces vectoriels topologiques, la composition des applications linéaires est hypocontinue de $\mathcal{L}(E_1, E_2) \times \mathcal{L}(E_2, E_3)$ dans $\mathcal{L}(E_1, E_3)$; lorsque les structures topologiques et bornologiques sont réparties différemment entre E_1, E_2 et E_3, on a une application bilinéaire *bornée :*

$$\mathcal{L}(E_1, E_2) \times \mathcal{L}(E_2, E_3) \to \mathcal{L}(E_1, E_3) ;$$

si E_1 et E_3 sont tous deux topologiques (resp. bornologiques), il faut imposer à E_2 d'être aussi topologique (resp. bornologique). Si F est un espace vectoriel topolo-

gique et si E est un espace vectoriel soit topologique, soit bornologique, l'application bilinéaire $(x, u) \mapsto u(x)$ de $E \times \mathcal{L}(E, F)$ dans F est hypocontinue ; lorsque E et F sont bornologiques, $E \times L(E, F) \to F$ est bornée.

Soit E et F des espaces vectoriels topologiques localement convexes ; on note $E \otimes_\pi F$ le produit tensoriel de E et F muni de la topologie localement convexe la plus fine rendant continue l'application bilinéaire canonique $E \times F \to E \otimes F$; la composition avec cette application induit alors un isomorphisme :

$$\mathcal{L}(E \otimes_\pi F, G) \to \mathcal{B}(E, F ; G),$$

pour tout e.l.c. G. Lorsque E et F sont des espaces normés, la topologie de $E \otimes_\pi F$ peut se définir au moyen de la *norme :*

$$\|z\|_1 = \inf \{\Sigma \|x_i\| . \|y_i\| | z = \Sigma x_i \otimes y_i\} ;$$

c'est visiblement une semi-norme : on démontre que c'est une norme en utilisant le théorème de Hahn-Banach, ce qui impose la restriction que le corps K soit égal à R, à C ou à un corps maximalement complet ; on a :

$$\|x \otimes y\|_1 = \|x\| . \|y\|,$$

pour $x \in E$ et $y \in F$. Si E et F sont séparés (resp. métrisables), il en est de même de $E \otimes_\pi F$. Le produit tensoriel $\otimes_\pi$, appelé produit tensoriel *projectif*, commute aux limites inductives finies. Pour E et F bornologiques de type convexe, on définit encore un produit tensoriel bornologique $E \otimes_\pi F$ avec la bornologie de type convexe la plus fine rendant $E \times F \to E \otimes F$ borné ; on a :

$$\mathcal{L}(E \otimes_\pi F, G) \simeq \mathcal{B}(E, F ; G)$$

pour tout G bornologique (ou topologique) de type convexe. Si E et F sont séparés, il en est de même de $E \otimes_\pi F$; le produit $\otimes_\pi$ bornologique commute aux limites inductives quelconques. Supposons maintenant E localement convexe et F bornologique de type convexe ; on munit $E \otimes F$ de la topologie localement convexe la plus fine qui rende hypocontinue $E \times F \to E \otimes F$, ce qui donne un e.l.c. $E \otimes_\pi F$, produit tensoriel *mixte* de E et F. On a, pour tout e.l.c. G, l'isomorphisme :

$$\mathcal{L}(E \otimes_\pi F, G) \simeq \mathcal{B}(E, F ; G),$$

où $\mathcal{B}(E, F ; G)$ est l'espace des applications bilinéaires hypocontinues ; il se peut que $E \otimes_\pi F$ ne soit pas séparé, même si E et F le sont ; le produit $\otimes_\pi$ mixte commute aux limites inductives quelconques.

Soit F un e.b.c. ; l'application $y \mapsto 1 \otimes y$ est une bijection de F sur $K \otimes F$ et permet de transporter à F la topologie du produit tensoriel mixte $K \otimes_\pi F$. On définit ainsi un e.l.c. ${}^tF$ dont un système fondamental de voisinages de 0 est formé par les disques de F qui absorbent tous les bornés ; la topologie de ${}^tF$ s'appelle la *topologie canonique* associée à la bornologie de F. Pour tout e.l.c. G, on a :

$$\mathcal{L}({}^tF, G) \simeq \mathcal{L}(K \otimes_\pi F, G) \simeq \mathcal{L}(F, \mathcal{L}(K, G)) \simeq \mathcal{L}(F, {}^bG) ;$$

en particulier, l'application identique $F \to {}^{bt}F$ (resp. ${}^{tb}G \to G$) est bornée (resp. continue) et on dit que F (resp. G) est *normal* si c'est un isomorphisme. On peut montrer que tout e.l.c. *métrisable* est normal.

Par complétion de $E \otimes_\pi F$, on obtient le *produit tensoriel complété* $E \widehat{\otimes}_\pi F$ (topologique, bornologique ou mixte selon le cas). Par exemple, si l'on désigne par l^1_F l'espace des suites (y_n) absolument sommables de vecteurs d'un espace de Banach F, muni de la norme $\|(y_n)\| = \Sigma \|y_n\|$, on a :

$$l^1 \widehat{\otimes}_\pi F \simeq l^1_F ;$$

plus généralement, si (X, μ) est un espace mesuré et F un espace de Banach, on a :

$$L^1(X, \mu) \hat{\otimes}_\pi F \simeq L^1_F(X, \mu),$$

et on en déduit que, si (Y. v) est un second espace mesuré, on a :

$$L^1(X, \mu) \hat{\otimes}_\pi L^1(Y, \nu) \simeq L^1(X \times Y, \mu \otimes \nu),$$

en utilisant le théorème de Fubini (cf. INTÉGRATION ET MESURE).

5. Espaces disqués

Sur un espace vectoriel E, on dit qu'une topologie $\mathcal{T}$ et une bornologie $\mathcal{B}$ vectorielles sont compatibles si elles satisfont les conditions suivantes :

(VB,) Tout voisinage de 0 pour $\mathcal{T}$ absorbe tout borné de $\mathcal{B}$;

(VB_2) L'adhérence (pour $\mathcal{T}$) d'un borné de $\mathcal{B}$ est encore bornée.

On appelle espace *disqué* un espace vectoriel muni d'une topologie localement convexe et d'une bornologie de type convexe compatibles. Par exemple, si E est un espace localement convexe, on obtient un espace disqué E^b en le munissant de sa bornologie canonique, qui est visiblement compatible avec la topologie (cf. début du chap. 4) ; on obtient un autre espace disqué E^p en prenant comme bornés les parties précompactes de E (dans le cas où K est localement compact). Si E est séparé, on en fait un espace disqué E^c en prenant comme système fondamental de bornés la famille des disques compacts ; une autre structure disquée E^s est définie par la bornologie vectorielle la plus fine sur E (cf. chap. 3, exemple 3).

L'espace disqué E est dit *séparé* si sa topologie l'est ; il est dit *quasi complet* si tout borné fermé est complet. Par exemple, lorsque E est un e.l.c. séparé, les espaces disqués E^c et E^s sont quasi complets ; si E est complet, E^b et Y sont quasi complets et l'on a $E^p = E^c$, mais les réciproques sont fausses. À tout espace disqué E on associe un quasi-complété $\hat{E}$ réunion des adhérences des bornés de E dans le complété $\hat{E}$ (pour la topologie) ; notons que. si E est un e.l.c., le quasi-complété de E^b n'est plus de la forme F^b pour un e.l.c. F.

Soit E et F des espaces disqués ; on note L(E, F) l'espace disqué dont les éléments sont les applications linéaires continues et bornées de E dans F, dont les bornés sont les ensembles à la fois équicontinus et équibornés d'applications linéaires et dont la topologie est celle de la convergence uniforme dans les bornés de E ; il est facile de vérifier que la bornologie et la topologie sont compatibles. Si F est séparé (resp. quasi complet), il en est de même de C(E, F). Pour trois espaces disqués E, F et G, on a encore un isomorphisme canonique :

$$\mathcal{L}(E, \mathcal{L}(F, G)) = \mathcal{L}(F, \mathcal{L}(E, G))$$

et on désigne par $\mathcal{B}(E, F ; G)$ l'espace correspondant d'applications bilinéaires de E X F dans G ; ses éléments sont les applications bornées et hypocontinues, de façon symétrique par rapport à E et F. Par exemple, la composition des applications est une application bornée et hypocontinue :

$$\mathcal{L}(E_1, E_2) \times \mathcal{L}(E_2, E_3) \rightarrow \mathcal{L}(E_1, E_3),$$

les espaces E_1, E_2 et E_3 étant disqués ; l'application $(x, u) \mapsto u(x)$ de E X C(E, F) dans F est bornée et hypocontinue, pour E et F disqués. Si E et F sont des e.l.c. séparés et G un espace disqué, toute application bilinéaire φ de E^s X F^s dans G est bornée ; si φ est hypocontmue, on dit qu'elle est *séparément continue.* Les applications bilinéaires que l'on rencontre en analyse sont

généralement hypocontinues pour des structures disquées convenables, mais rarement continues.

Le théorème de Banach-Steinhaus est lié à la structure des espaces C(E, F) et $\mathcal{B}$(E, F ; G). Il s'énonce ainsi (cf. espaces vectoriels NORMÉS, chap. 4) :

Théorème. a) Soit E et F des espaces disqués ; on suppose que la topologie de E est une *topologie* **de** *Baire.* Alors tout ensemble équiborné d'applications linéaires continues de E dans F est équicontinu.

b) Soit E, F et G des espaces disqués ; on suppose que E et F sont *métrisables* et que l'un d'eux est *complet.* Alors tout ensemble équihypocontinu d'applications bilinéaires de E X F dans G est équicontinu.

De (*a*) on déduit par exemple que, si une suite (u_n) d'applications linéaires continues de E dans F converge simplement vers une limite u, cette limite est encore linéaire continue ; car l'ensemble $\{u_n\}$ est simplement borné, donc équicontinu et, par suite, son adhérence $\{u_n\}_n \cup \{u\}$ est encore équicontinue.

La propriété de Baire n'est pas nécessaire pour la conclusion de (*a*). On dit qu'un espace disqué E est *infratonnelé* si, pour tout espace disqué F, toute partie équibornée de C(E, F) est équicontinue ; il revient au même de dire que tout disque fermé de E qui absorbe les bornés est un voisinage de 0. Pour qu'un espace disqué E soit infratonnelé, il faut et il suffit que toute application linéaire bornée de E dans un espace de Banach soit continue dès que son graphe est fermé. Si E et F sont des espaces disqués métrisables dont l'un est infratonnelé, tout ensemble équihypocontinu d'applications bilinéaires de E X F dans un espace G est équicontinu. Tout quotient d'un espace infratonnelé est infratonnelé ; la somme et le produit d'une famille d'espaces infratonnelés sont infratonnelés. On dit qu'un espace localement convexe E est *tonnelé* (resp. *quasi-tonnelé*) si E^s (resp. E^b) est infratonnelé ; les espaces de Fréchet sont tonnelés et les e.l.c. métrisables sont quasi-tonnelés.

Le produit tensoriel E ⊗ F de deux espaces disqués est muni de la structure disquée (topologie et bornologie) la plus fine rendant bornée et hypocontinue l'application bilinéaire canonique E × F → E ⊗ F ; on le note $E \otimes_\pi F$ et on a :

$$\mathcal{L}(E \otimes_\pi F, G) \rightleftarrows \mathcal{B}(E, F ; G)$$

pour tout espace disqué G. Le produit tensoriel disqué $\otimes_\pi$ commute aux limites inductives ; si E et F sont séparés (resp. infratonnelés), il en est de même de $E \otimes_\pi F$. On définit ainsi différentes structures disquées sur le produit tensoriel de deux e.l.c. E et F séparés : le produit tensoriel « inductif » :

$$E \otimes_{\pi\sigma} F = E^s \otimes_\pi F^s$$

et les deux structures :

$$E \otimes_{\pi\gamma} F = E^c \otimes_\pi F^c,$$
$$E \otimes_{\pi\beta} F = E^b \otimes_\pi F^b ;$$

lorsque E et F sont tonnelés, les topologies sous-jacentes à ces trois structures coïncident et font de E ⊗ F un e.l.c. tonnelé. Les trois topologies coïncident encore si les espaces E et F sont métrisables, l'un d'eux étant tonnelé : ces topologies sont alors identiques à celle du produit tensoriel topologique $E \otimes_\pi F$.

6. Dualité

Soit E un espace vectoriel topologique, bornologique ou disqué ; son *dual* est, par définition, l'espace C(E, K) = E' des for-

mes linéaires sur E. Ainsi, le dual d'un espace vectoriel topologique (resp. bornologique, disqué) est un espace vectoriel bornologique (resp. topologique, disqué). Par exemple, on définit l'espace C'(X) des distributions à support compact dans une variété différentiable X, l'espace S' des distributions tempérées dans $\mathbf{R}^n$ et l'espace 0'(U) des fonctionnelles analytiques dans l'ouvert U de $\mathbf{C}^n$ comme les duaux respectifs des e.l.c. donnés dans les exemples 4, 5 et 6 du chapitre 1 ; l'espace 'D'(X) des distributions dans X est le dual de l'exemple 4 du chapitre 3 et l'espace $\mathcal{O}'(U)$ des fonctionnelles analytiques dans U peut encore être défini comme le dual de l'exemple 5 du chapitre 3 (cf. DISTRIBUTIONS). Si E est un e.b.c., on a :

$$({}^tE)' = \mathcal{L}({}^tE, K) \simeq \mathcal{L}(E, {}^bK) \simeq {}^b(E');$$

en particulier, si F est un e.l.c. normal (cf. chap. 4), par exemple métrisable, on a E' = $^b((^bE)')$. Le dual d'un e.l.c. séparé E admet diverses structures disquées, avec toujours la même bornologie, dont les bornés sont les ensembles équicontinus de formes linéaires : le *dual faible* E', = (E')', muni de la topologie de la convergence simple dans E ; l'espace E', = $(E^c)'$, muni de la topologie de la convergence uniforme dans les disques compacts de E ; le *dual fort* E'_b = $(E^b)'$, muni de la topologie de la convergence uniforme dans les parties bornées de E. Si E est un e.l.c. normal, on a l'égalité E'_b = $((^bE)')^b$; les duaux forts d'espaces de Fréchet ont des applications importantes en analyse complexe et ont été systématiquement étudiés par A. Grothendieck.

Le bidual E" = (E')' d'un espace E a une structure de même nature que celle de E et on a un morphisme de bidualité E ⟶ E" (cf. algèbre LINÉAIRE ET MULTILINÉAIRE) ; ce morphisme est strict. Lorsque E est un e.l.c. ou un espace disqué séparé, le morphisme de bidualité est *injectif* ; il résulte en effet du théorème de Hahn-Banach (cf. CONVEXITÉ) que, pour tout élément $x \neq 0$ de E, il existe un $x' \in$ E' tel que x'(x) $\neq$ 0 (prolonger à E une forme linéaire non nulle définie sur la droite Kx). Il n'en est plus de même pour un e.b.c. E, car il se peut qu'on ait E' = 0 même si E est séparé et non nul ; pour que E ⟶ E" soit injectif, il faut et il suffit que E soit séparé et que sa bornologie soit compatible avec la topologie canonique associée, c'est-à-dire que l'adhérence dans 'E d'un borné de E soit encore bornée : on dit alors que E est *régulier*. Tout e.b.c. séparé normal est régulier ; toute limite projective et toute somme d'e.b.c. réguliers sont régulières. Si E et F sont des e.l.c. et si F est séparé, alors $\mathcal{L}$(E, F) est régulier ; il en est de même si E est un e.l.c. ou un e.b.c. et si F est un e.b.c. régulier.

On dit que E est *réflexif* si le morphisme de bidualité est un isomorphisme E ⥲ E". Par exemple, soit (X, μ) un espace mesuré etp un nombre réel > 1 ; l'espace L^p (X, μ) des classes de fonctions de puissancep-ième intégrable (cf. chap. 1, exemple 2) est réflexif, car son dual est isomorphe à L^q(X, μ), avec $1/p + 1/q = 1$ (cf. INTÉGRATION ET ME~URE). Sur un corps ultramétrique K, les seuls espaces normés réflexifs sont les espaces de dimension finie. On dit qu'un espace localement convexe E est un *espace de Schwartz* si, pour toute application linéaire continue f de E dans un espace de Banach, il existe un voisinage U de 0 dans E tel que f (U) soit relativement compact ; ainsi. l'espace C(X) de l'exemple 4 du chapitre 1, l'espace $\mathcal{O}$(U) de l'exemple 6 du chapitre 1 et l'espace 'd>'(X) des distributions (cf. *supra)* sont des espaces de Schwartz : cela résulte du théorème des accroissements finis et du théorème

d'Ascoli ; on peut montrer que les espaces de Schwartz complets sont réflexifs. Pour tout e.l.c. séparé E, les espaces disqués E^s et E^c sont réflexifs ; si E est réflexif, il en est de même de E^b ; mais la réciproque n'est pas vraie en général.

La théorie de la dualité se développe à l'aide de la notion de *système dual;* on appelle ainsi un triplet (E, F, u) où E et F sont des espaces vectoriels et où u : $E \times F \rightarrow K$ est une forme bilinéaire. Pour toute partie X de E, on appelle *polaire* de X l'ensemble :

$$X^0 = \{y \in F \mid \forall x \in X, |u(x,y)| \leqslant 1\};$$

de même, on définit le polaire $Y^0 \subset E$ d'une partie Y de F. Il est clair que X^0 est un disque et que l'application $X \rightarrow X^0$ est décroissante. De plus. on a :

$$X \subset X^{00},$$
$$X^0 = X^{000},$$

et $(\lambda X)^0 = (1/\lambda)X^0$, pour $A \in K^*$; enfin, pour toute famille (X,) de parties de E, on a :

$$\left(\bigcup X_i\right)^0 = \bigcap X_i^0, \quad \left(\bigcap X_i\right)^0 \supset \bigcup X_i^0.$$

Si l'on a X C E et Y C F, pour que X^0 absorbe Y, il faut et il suffit que Y^0 absorbe X. Considérons un autre système dual (E,, F_1, u_1) et des applications linéairesfde E dans E_1 et g de F, dans F telles que :

$$u_1(f(x), y_1) = u(x, g(y_1)),$$

pour $x \in E$ et $y_1 \in F_1$; si on a $X \subset E$, il vient $(f(X))^0 = g^{-1}(X^0)$.

Soit (E, F, u) un système dual ; on note ${}^\sigma E$ l'espace E muni de la *topologie faible* dont un système fondamental de voisinages de 0 est formé par les polaires de parties finies de F (topologie de la convergence simple dans F). L'adhérence de 0 dans ${}^\sigma E$ est F^0 et le dual de ${}^\sigma E$ s'identifie, au moyen de u, à F/E^0 muni de la bornologie vectorielle la plus fine. Le *bipolaire* X^{00} d'une partie X de E est son enveloppe disquée faiblement fermée, c'est-a-dire fermée dans ${}^\sigma E$; ce résultat, connu *sous* le nom de *théorème des bipolaires, se* démontre à partir du théorème de Hahn-Banach. Les parties faiblement bornées de E (bornées de ${}^{b\sigma}E$) sont exactement les parties *faiblement précompactes.* Appliquons cela au système dual (F', F, u) où F est un e.l.c. et où $u(x', x) = x'(x)$ pour $x \in F$ et $x' \in F'$: la topologie faible de F' est celle de F'_s et ses parties équicontinues sont faiblement bornées ; or les disques équicontinus sont les polaires dans le dual algébrique F* des voisinages de 0 dans F, ce qui montre qu'ils sont complets pour la topologie faible, car ce sont des fermés dans F* qui est complet. On démontre ainsi le théorème suivant, qui généralise le « principe de choix » de Hilbert (théorème d'Alaoglu pour les espaces normés) :

Théorème. Soit F un e.l.c. : les parties équicontinues de F' sont faiblement relativement compactes, c'est-à-dire d'adhérence compacte dans F',.

Voici un autre résultat fondamental de la théorie :

Théorème de Mackey. Soit (E, F, u) un système dual ; on suppose E et F munis de bornologies de type convexe compatibles avec les topologies faibles et telles que u soit borné et on munit E (resp. F) de la topologie dont un système fondamental **de** voisinages de 0 est formé des polaires de bornés de F (resp. E). Si les bornés de E et de F sont précompacts pour les topologies précédentes (ce qui revient à dire que l'application u est faiblement continue sur le produit d'un borné de E par un borné de F), le dual de E s'identifie, au moyen de u, au quasi-complété de F.

Supposons maintenant F faiblement séparé (soit $E^0 = 0$) ; on appelle *topologie de Mackey* sur E la topologie dont un système fondamental de voisinages de 0 est formé des polaires de disques faiblement compacts de F. Il résulte du théorème de Mackey que la topologie de Mackey sur E est la plus fine des topologies pour lesquelles toute forme linéaire continue peut s'écrire $x \mapsto u(x, y)$ pour un $y \in F$.

CHRISTIAN HOUZEL

Bibliographie

S. BANACH, *Théorie des opérations linéaires,* Varsovie, 1932, plusieurs fois réédité / N. BOURBAKI, *Espaces vectoriels topologiques*, chap. I à V, nouvelle version, Masson, Paris, 1981 / A. GROTHENDIECK, *Espaces vectoriels topologiques*, Publicação de Sociedade de Mathemática de São Paulo, 2e éd., 1958 / C. HOUZEL, *Séminaire Banach*, Springer, Berlin, 1972.

TRANSCENDANTS NOMBRES

Si la notion de nombre irrationnel remonte aux Grecs, l'idée de nombre transcendant n'a pu se dégager qu'après la création de notations algébriques assez développées pour que le concept de polynôme de degré quelconque puisse être clairement formulé ; aussi est-ce seulement au XVIIe siècle que l'on commence à faire la distinction entre les nombres *algébriques,* tels $\sqrt{3/2}$ ou $\cos(\pi/n)$ pour n entier, qui sont racines de polynômes à coefficients *entiers,* les autres nombres réels étant qualifiés de *transcendants.* L'existence de nombres transcendants n'a été prouvée qu'au XIXe siècle ; s'il est facile de construire des nombres transcendants, la question de savoir si un nombre donné est ou non transcendant est généralement un problème fort difficile. L'exemple le plus célèbre est celui du nombre π, dont la transcendance n'a été démontrée qu'en 1882 ; ce résultat prouvait définitivement l'impossibilité de la « quadrature du cercle », c'est-à-dire le problème, posé depuis les Grecs, de la construction géométrique « par la règle et le compas » d'une longueur égale à la circonférence de diamètre unité ; il est facile, en effet, de montrer qu'une telle construction ne peut jamais donner que des longueurs dont la mesure est un nombre algébrique (et même un nombre algébrique d'un type très particulier).



L'existence des nombres transcendants

Il est commode d'étendre la définition des nombres algébriques aux nombres complexes, et d'appeler encore nombre transcendant un nombre complexe non algébrique. J. Liouville a établi, en 1844, l'existence des nombres transcendants par une construction fondée sur la propriété, découverte par lui, de « mauvaise approximation » des nombres irrationnels algébriques par les nombres rationnels (cf. approximations DIOPHANTIENNES). En 1873, G. Cantor déduisit l'existence des nombres transcendants de son théorème prouvant que l'ensemble de tous les nombres réels est non dénombrable : il suffit, en effet, de prouver que l'ensemble A de tous les nombres algébriques est dénombrable. Pour cela, associons à chaque polynôme à coefficients entiers :

$$P(X) = a_0X^n + a_1X^{n-1} + \cdots + a_n,$$

sa *hauteur :*

$$\mathrm{ht}(P) = \max_j |a_j|,$$

Comme un polynôme n'a qu'un nombre fini de racines, l'ensemble A_N des nombres algébriques qui sont racines de polynômes à coefficients entiers de degré $\leqslant$ N et de hauteur $\leqslant$ N est un ensemble *fini* ; comme A est réunion des A_N, pour N = 1, 2, ... l'ensemble A est dénombrable.

L'ensemble des nombres transcendants n'est donc pas dénombrable ; en termes imagés, on **peut** dire qu'un nombre pris au hasard (par exemple en se donnant au hasard son développement décimal illimité) n'a « aucune chance » d'être algébrique.

Valeurs transcendantes de fonctions entières

Le premier résultat profond sur les nombres transcendants fut obtenu par C. Hermite en 1872 : par une méthode très originale reposant sur l'approximation de la fonction exponentielle e^z par des fonctions rationnelles, il put montrer que le nombre e est transcendant, et **c'est** par une extension de la méthode d'Hermite que Ferdinand von Lindemann, en 1882, prouva que π est aussi transcendant. De nouveaux résultats de cette nature n'apparurent qu'après 1929 ; ils concernent, comme les précédents, des nombres qui sont des valeurs prises par certaines fonctions entières ou méromorphes (ou leurs fonctions inverses) pour des valeurs algébriques de la variable ; les méthodes, développées à partir d'idées de C. L. Siegel et de A. Gelfond, raffinées par T. Schneider et récemment par A. Baker, utilisent, comme celle d'Hermite, des propriétés des fonctions entières d'une variable complexe. Aucune méthode n'a encore été trouvée pour des nombres qui ne sont pas donnés de cette manière, par exemple la constante d'Euler :

$$\gamma = \lim_{n\to\infty}\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} - \operatorname{Log} n\right),$$

dont on ne sait même pas si elle est irrationnelle (cf. calculs ASYMPTOTIQUES chap. 2).

Les théorèmes d'Hermite et de Lindemann sont des cas particuliers du résultat suivant.

Théorème I. Un nombre complexe $\alpha \neq 0$ ne peut être tel que α et e^{α} soient tous deux algébriques.

En effet, ce théorème entraîne, en faisant $\alpha = 1$, que e est transcendant et, en faisant $\alpha = i\pi$, que π est transcendant.

La méthode de Siegel-Gelfond-Schneider déduit le théorème I d'un théorème plus général concernant les valeurs de deux fonctions entières liées par des *équations différentielles algébriques* :

Théorème II. Soit f et g deux fonctions entières d'ordre fini ρ, c'est-à-dire vérifiant des majorations,

$$(1) \qquad \begin{aligned} |f(z)| &\leqslant \exp(a|z|^{\rho}), \\ |g(z)| &\leqslant \exp(a|z|^{\rho}), \end{aligned}$$

pour **tout** $z \in \mathbb{C}$. On suppose que :

1° Les fonctions f et g sont algébriquement indépendantes, c'est-à-dire qu'il n'existe aucun polynôme P(X, Y) non nul à coefficients complexes tel que $P(f(z), g(z)) \equiv 0$.

2° On a des relations différentielles :

$$(2) \qquad \begin{aligned} f'(z) &= Q(f(z), g(z)), \\ g'(z) &= R(f(z), g(z)), \end{aligned}$$

où Q et R sont des polynômes à coefficients dans un corps de nombres algébriques K de degré fini d sur le corps des rationnels $\mathbb{Q}$.

Supposons alors qu'il y ait m nombres complexes distincts $w_1, \dots, w_m$ tels que

les $2m$ nombres $f(w_j)$ et $g(w_j)$ appartiennent à K. Cela n'est possible que si $m \leqslant 10\ pd$.

Pour en tirer le théorème 1, on prend f(z) $=$ z et g(z) $= e^z$, qui vérifient les conditions (1) et (2) ; si a et e^{α} étaient algébriques, ils appartiendraient à un même corps de nombres K dc degré fini ; mais alors tous les nombres $n\alpha$ et $e^{n\alpha}$ appartiendraient aussi à K pour tout entier n, ce qui contredit le théorème II.

Si β est un nombre algébrique irrationnel, les fonctions e^t et $e^{\beta t}$ vérifient les conditions du théorème II en prenant pour K un corps contenant β. On en déduit le *théorème de Gelfond-Schneider* :

Théorème III. Si a est un nombre algébrique autre que 0 et 1 et β un nombre algébrique irrationnel, le nombre $\alpha\beta$ = exp (β Log *a)*, avec une détermination quelconque du logarithme, est transcendant.

En effet, dans le cas contraire, les valeurs de e^t et $e^{\beta t}$ pour tous les nombres n Log a, avec n entier quelconque, appartiendraient à un corps K' contenant a et α^{β}, contrairement au théorème II.

Par exemple, le nombre $2^{\sqrt{2}}$ et le nombre e^{π}, qui est une détermination de (- $1)^{-i}$, sont transcendants.

L'idée de la démonstration du théorème II est de former une suite (F,(z)) de fonctions entières d'ordre ρ, où l'indice s prendra des valeurs entières *arbitrairement grandes*, qui possèdent les propriétés suivantes, où C,, C_2, sont des constantes indépendantes de s,

(*a*) $\quad |F_s(z)| \leqslant s^{3s} \exp(C_1 s^{1/2} |z|^{\rho}),$

(*b*) $\left(\dfrac{d}{dz}\right)^k F_s(w_j) = 0,$

$$0 \leqslant k < s, \quad 1 \leqslant j \leqslant m.$$

(c) Il existe un w_j, par exemple w_1, tel que :

$$\left(\frac{d}{dz}\right)^s F_s(w_1) = \gamma \neq 0,$$

et tel que :

$$(3) \qquad |\gamma^{-1}| \leqslant C_2 s^{5(d-1)s}.$$

On remarque alors que la fonction :

$$G(z) = \frac{F_s(z)}{(z - w_1)^s \dots (z - w_m)^s}$$

est une fonction *entière* en vertu de *(b)* qui, pour z $\leqslant$ R, où R est assez grand, vérifie, en vertu de *(a)* et du principe du maximum, l'inégalité :

$$(4) \qquad G(z) \leqslant R^{-ms} s^{3s} \exp(C_2 s^{1/2} R^{\rho}).$$

D'autre part, on a :

$$\gamma = \left(\frac{d}{dz}\right)^s F_s(w_1) = s!\, G(w_1) \prod_{j=2}^{m} (w_1 - w_j)^s.$$

On prend R = $s^{1/(2\rho)}$, et on déduit de (4) la majoration :

$$(5) \qquad |\gamma| \leqslant C_3^s s^{(8\rho - m)s/(2\rho)}$$

et, en comparant avec la minoration de y donnée par (3), on obtient aisément $m \leqslant 10\ pd$.

Tout revient donc à satisfaire aux conditions *(a)*, *(b)* et (c). On se donne un entier r arbitrairement grand divisible par $2m$ et on pose $n = r^2/(2\ m)$. On considère la fonction :

$$F(z) = \sum_{\mu, \nu = 1} b_{\mu\nu} f(z)^{\mu} g(z)^{\nu},$$

où l'on détermine les coefficients $b_{\mu\nu}$ *dans* K tels que :

$$(6) \quad \left(\frac{d}{dz}\right)^k F(w_j) = 0,$$

$$0 \leqslant k < n, \quad 1 \leqslant j \leqslant m.$$

C'est un système de mn équations linéaires à $r^2 = 2\,mn$ inconnues $b_{\mu\nu}$, dont les coefficients sont par hypothèse dans K, en vertu des relations (2) et de l'hypothèse sur les $f(w_j)$ et $g(w_j)$. On montre par des majorations élémentaires que l'on peut prendre pour solutions $b_{\mu\nu}$ de ce système des *entiers de* K non tous nuls tels que :

(7) $$|b_{\mu\nu}| \leqslant C_4 n^{2n}.$$

L'hypothèse que f et g sont algébriquement indépendantes implique alors que F n'est pas identiquement nulle. Il y a donc un plus petit entiers tel que F = F_s vérifie (b) et que :

$$\left(\frac{d}{dz}\right)^s F(w_j)$$

ne soit pas nul pour un j au moins ; en vertu de (6), on a nécessairement $s \geqslant n$. On prouve élémentairement qu'il y a un nombre rationnel $c \leqslant C_5^s$ tel que $c\gamma$ soit un *entier* algébrique de K et que $c\gamma$ et tous ses conjugués soient majorés par $C_6\, s^{5s}$ en valeur absolue. Pour la *norme* de $c\gamma$, qui est un entier rationne1 non nul, on a alors les inégalités :

$$1 \leqslant |N_{K/Q}(c\gamma)| \leqslant C_7 s^{5(d-1)s} |\gamma|,$$

ce qui donne (3) ; quant à la majoration (u) de F_s, elle se déduit aisément des majorations defet g et de (7).

A. Baker a généralisé les théorèmes I et III ; désignons par L l'ensemble des nombres complexes z tels que e^z soit un nombre algébrique ; c'est évidemment un sous-espace vectoriel de C sur le corps Q des nombres rationnels. Alors, on a les résultats suivants de Baker :

Théorème IV. (a) Si $z_1, \ldots z_n$ sont des éléments de L linéairement indépendants sur $\mathbf{Q}$, ils sont aussi linéairement indépendants sur le corps de tous les nombres algébriques, autrement dit il ne peut exister de relation :

$$\beta_1 z_1 + \ldots + \beta_n z_n = 0,$$

avec des coefficients algébriques $\beta_1, \ldots \beta_n$ non tous nuls.

(h) Si $z_1, \ldots z_n$ sont des éléments de L, il ne peut exister de relation :

$$\beta_0 = \beta_1 z_1 + \ldots + \beta_n z_n,$$

où $\beta_0, \beta_1, \ldots \beta_n$ sont des nombres algébriques et où l'on a $\beta_0 \neq 0$.

On déduit aussitôt du théorème IV que, si $\beta_1, \ldots \beta_n$ sont des nombres algébriques tels que $1, \beta_1, \ldots \beta_n$ soient linéairement indépendants sur $\mathbf{Q}$, alors le nombre :

$$\alpha_1^{\beta_1}\alpha_1^{\beta_2} \quad \alpha_n^{\beta_n}$$

est transcendant pour tous les α_j algébriques différents de 0 et de 1. De même, si $\alpha_1, \ldots \alpha_n, \beta_0, \beta_1, \ldots \beta_n$ sont des nombres algébriques non nuls, le nombre :

$$e^{\beta_0}\alpha_1^{\beta_1} \ldots a_n^{\beta_n}$$

est transcendant. On déduit aussi par exemple du théorème IV que π + Log α est transcendant pour tout nombre algébrique $\alpha \neq 0$.

La méthode de démonstration du théorème IV est une extension de celle du théorème II, mais utilisant une technique plus subtile ; pour (h), par exemple, on raisonne par l'absurde en supposant l'existence d'une relation :

$$z_n = \beta_0 + \beta_1 z_1 + \ldots + \beta_{n-1} z_{n-1},$$

et on considère la fonction entière de n variables complexes $t_0, t_1, \ldots t_{n-1}$:

$$\Phi(t_0, \ldots, t_{n-1}) = \sum_{\lambda} p(\lambda_0, \ldots, \lambda_n) t_0^{\lambda_0} \exp u(t_0, \ldots, t_n),$$

où :

$$u(t_0, t_1, ..., t_n) = \lambda_n \beta_0 t_0 + (\lambda_1 + \lambda_n \beta_1) z_1 t_1 + ... + (\lambda_{n-1} + \lambda_n \beta_{n-1}) z_{n-1} t_{n-1},$$

les λ_n étant des entiers tels que $0 \leqslant$ A, $\leqslant c_1$. On impose aux $p(\lambda)$ d'être entiers et choisis de sorte que Φ et toutes ses dérivées partielles jusqu'à un ordre c_2 s'annulent lorsqu'on y fait $t_1 = t_2 = ___ = t_{n-1} = w$ et que w est un entier tel que $0 \leqslant w \leqslant c_3$. Par un choix convenable des constantes c_1, c_2, c_3 et l'emploi de majorations tirées de la théorie des fonctions holomorphes, on aboutit à une contradiction.

Indépendance algébrique de nombres transcendants

La transcendance d'un nombre a signifie qu'il n'est pas racine d'un polynôme à coefficients entiers. Plus généralement, n nombres complexes a,, α_2, ... a,, sont dits *algébriquement indépendants* s'il n'existe aucun polynôme non nul P(T_1, ... T,,) à n indéterminées et à coefficients entiers tel que P(α_1, ... a,,), = 0, ce qui implique bien entendu que a,, ..., *a,,* sont transcendants. On n'a que peu de résultats sur cette question ; par exemple, on ignore si e et π sont algébriquement indépendants. On conjecture que, sous les conditions (a) du théorème IV, les z_j sont algébriquement indépendants.

Les résultats positifs les plus intéressants sont les suivants :

Théorème V (Lindemann). Si a,, ..., a,, sont des nombres algébriques linéairement indépendants sur $\mathbf{Q}$, les nombres e^{α_1}, ... e^{α_n} sont algébriquement indépendants.

Théorème VI (Siegel). Si J',, est la fonction de Bessel d'indice 0, alors les nombres J,,(a) et $J'_0(\alpha)$ sont algébriquement indépendants pour tout nombre algébrique *a ≠ 0.*

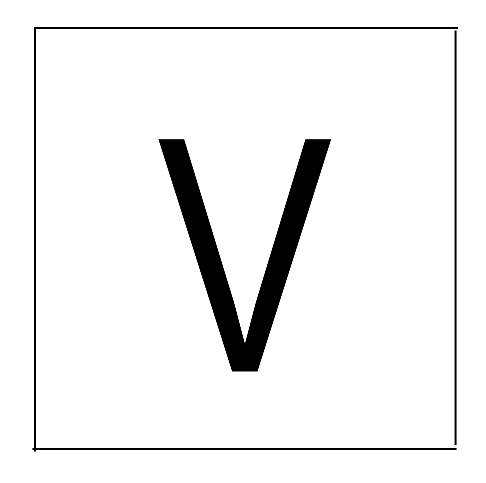
Une même méthode, due à Siegel, permet de démontrer ces deux résultats. Elle applique une idée analogue à celle de la démonstration du théorème II, utilisant le fait que les fonctions e^{az} et $J_0(z)$ vérifient une équation différentielle *linéaire* homogène et des majorations tirées de la théorie des fonctions analytiques ; mais les détails de la démonstration sont beaucoup plus délicats et compliqués.

JEAN DIEUDONNÉ

Bibliographie

A. BAKER, *Transcendental Number Theory*, Cambridge Univ. Press, Cambridge, 1990 / D. BERTRAND et M. WALDSCHMIDT éd., *Approximations diophantiennes et nombres* transcendants. Colloque de Luming 1982, Birkauser, Stuttgart, 1983 ; *Fonctions abéliennes et nombres transcendants*, Société mathématique de France. Paris, 1980 / D. BERTRAND et al., Les *Nombres transcendants*, Gauthier-Villars, Paris, 1984 / G.V. CHUDNOVSKY, *Contributions to the Theory of Transcendantal Numbers*, American Mathematical Society, Providence (R.I.), 1984 / A. B. SHIDLOVSKII, *Transcendantal Numbers*, De Gruyter, Hawthorne (N.Y.), 1989.

TRIGONOMÉTRIQUES SÉRIES → SÉRIES TRIGONOMÉTRIQUES



VARIATIONS CALCUL DES

L'étude d'une fonction à valeurs réelles comporte en particulier la détermination de ses extrémums. C'est là un des objets du calcul différentiel classique lorsque la source de cette fonction est un espace numérique ; c'est l'objet de ce qu'Euler a appelé le *calcul des variations* lorsque cette source est un espace fonctionnel.

On rencontre déjà dans la plus haute antiquité des problèmes d'une telle nature. La légende ne veut-elle pas que Didon, lorsqu'elle fonda Carthage, ait délimité la plus grande étendue qu'elle pût circonscrire à l'aide de lanières découpées dans la peau d'un taureau? Et il est bien connu que les Grecs caractérisaient un segment de droite comme la ligne de plus petite longueur joignant ses extrémités.

Ce n'est cependant qu'au XVIIIe siècle, à la suite de l'essor du calcul infinitésimal, qu'Euler et Lagrange établirent les fondements du calcul des variations et donnèrent une première condition d'extrémum. Cette équation d'Euler-Lagrange allait jouer un rôle très important, surtout en physique, où elle justifiait les principes variationnels : principe *de Fermat* pour la propagation de la lumière dans les milieux différemment réfringents ; *principes de moindre action* de Maupertuis et Hamilton pour la détermination des mouvements en mécanique analytique.

La recherche de conditions d'extrémum se poursuivit aux XVIIIe et XIXe siècles, notamment avec les travaux de Legendre, Jacobi et Weierstrass, pour aboutir au début du XXe siècle à une théorie bien élaborée que l'on situe aujourd'hui dans le cadre du calcul différentiel au sens de Fréchet dans les *espaces de Banach*. Mais de difficiles problèmes relatifs à l'existence de ces extrémums restent encore ouverts.

Plus récemment, les travaux de Morse relancèrent l'intérêt porté au calcul des variations. Utilisant à la fois des techniques d'analyse fonctionnelle, de topologie algébrique et de topologie différentielle, ils sont à l'origine de ce qu'on appelle maintenant l'analyse différentielle globale, une des théories carrefours de la mathématique actuelle.

Il faut enfin mentionner le contrôle optimal, terminologie d'origine anglo-saxonne fréquemment remplacée par « commande optimale ». Par ses problèmes d'optimisation de fonctionnelles sur des espaces de solutions d'équations différentielles avec paramètres de contrôle, il s'intègre en effet au calcul de variations. Mais la recherche de solutions qui peuvent être discontinues y conduit au développement de techniques fort différentes ; on n'en parlera pas ici.

Quelques problèmes classiques

La brachistochrone

On considère dans le champ de la pesanteur deux points A et B et un point matériel M se déplaçant sans frottement sur une courbe d'extrémités A et B. Déterminer la courbe. appelée brachistochrone, pour laquelle le temps de parcours est minimal lorsque le point M part du point A avec une vitesse nulle.

Ce problème, dont la solution est en général un arc de cycloïde, avait déjà été considéré par Galilée, qui avait remarqué que ce minimum n'était pas réalisé par le segment de droite. Résolu en 1697, en particulier par Jean Bernoulli, Jacques Bernoulli et Newton, il allait attirer l'attention des mathématiciens de l'époque sur les problèmes variationnels.

En admettant que sa solution soit une courbe plane ayant une équation de la forme $y = f(x)$, on peut en donner la formulation analytique suivante : Déterminer la fonction continûment dérivable $y = f(x)$ vérifiant les conditions $f(x_0) = y_0$ et $f(x_1) = y_1$ qui minimise l'intégrale :

$$J = \int_{x_0}^{x_1} \frac{\sqrt{1+y'^2}}{\sqrt{2g(y_0 - y)}} dx.$$

La surface minimale de révolution

Étant donné dans un plan P un axe 3 et deux points A et B situés d'un même côté de A, déterminer la courbe du plan P, d'extrémités A et B, engendrant par révolution autour de Δ une surface dont l'aire est minimale. Sous des hypothèses analogues à celles qui ont été faites précédemment, on est ici amené à minimiser l'intégrale :

$$J = 2\pi \int_{x_0}^{x_1} y\sqrt{1+y'^2}\, dx.$$

La solution est en général un arc de chaînette :

$$y = \lambda \operatorname{ch} \frac{x+\mu}{\lambda},$$

où ch désigne le cosinus hyperbolique.

Les géodésiques

Étant donné une surface S dans $\mathbf{R}^3$ et deux points A et B de S, déterminer les courbes tracées sur S d'extrémités A et B et de longueur minimale.

De ce point de vue, le problème de la distance d'un point à une courbe suggère, d'ailleurs, la généralisation des exemples précédents à la considération de courbes dont les extrémités sont non pas fixées mais mobiles sur deux courbes données (problèmes variationnels à extrémités variables).

Le problème isopérimétrique

Déterminer parmi les courbes planes fermées sans point double de longueur donnée celle dont l'intérieur a la plus grande surface : c'est le problème de Didon.

Ce problème, dont la solution est le cercle, est un problème d'extrémum lié. Il fut résolu par Jacques Bernoulli en 1697 et joua également un rôle important dans l'essor du calcul des variations.

Le problème de Plateau

Les quatre exemples précédents concernent des espaces de courbes : on dit que ce sont des problèmes variationnels de dimension 1. On peut bien évidemment concevoir des problèmes de dimensions supérieures. Le plus célèbre d'entre eux est celui du physicien belge J. Plateau : Étant

donné dans l'espace une courbe fermée sans point double, déterminer une surface d'aire minimale ayant cette courbe pour bord. Ce problème est essentiel dans l'étude des lames minces liquides.

Présentation analytique d'un problème variationnel

À la lumière des exemples qui viennent d'être présentés, on peut donner la formulation suivante d'un problème variationnel simple de dimension 1 à extrémités fixes.

Soit $\mathcal{D}$ l'espace affine des fonctions f à valeurs réelles continûment dérivables sur l'intervalle $[a, b]$ et vérifiant $f(a) = \alpha$ et $f(b) = \beta$. L'espace vectoriel $\mathcal{E}$ associé à cet espace affine peut s'interpréter comme l'espace des fonctions ω continûment dérivables sur $[a, b]$ et vérifiant $\omega(a) = \omega(b) = 0$.

On munit l'espace $\mathcal{D}$ des deux topologies $\mathcal{C}^0$ et $\mathcal{C}^1$ définies respectivement par les normes de la convergence uniforme :

$$\|f\| = \sup_{[a,b]} |f(x)| ,$$
$$\|f\|_1 = \sup_{[a,b]} (|f(x)| + |f'(x)|).$$

La topologie $\mathcal{C}^1$ est plus fine que la topologie $\mathcal{C}^0$.

Soit $F(x, y, y')$ une fonction à valeurs réelles deux fois continûment différentiable sur l'espace $[a, b] \times \mathbf{R} \times \mathbf{R}$. On peut lui associer la fonctionnelle J sur $\mathcal{D}$ déterminée par :

$$f \mapsto \int_a^b F(x, f(x), f'(x))\, dx.$$

On dit alors qu'une fonction f de $\mathcal{D}$ est une solution du problème variationnel correspondant à la fonction F si elle est un minimum de J sur $\mathcal{D}$, c'est-à-dire si l'on a $J(f) \leqslant J(g)$ pour **tout** $g \in \mathcal{D}$. On dit également que f est un minimum relatif faible (resp. fort) de J s'il existe $\varepsilon > 0$ tel que l'on ait $J(f) \leqslant J(g)$ pour tout $g \in \mathcal{D}$ vérifiant $\|f-g\|_1 \leqslant \varepsilon$ (resp. $\|f-g\| \leqslant \varepsilon$).

Naturellement un minimum au sens fort est également un minimum au sens faible. Mais cette distinction se trouve justifiée par le fait que la fonction J, qui est continue pour la topologie $\mathcal{C}^1$, ne l'est pas en général pour la topologie $\mathcal{C}^0$.

On peut maintenant, avec J. L. Lagrange, considérer un élément ω de $\mathcal{E}$ comme **une** « variation » de la fonction f de $\mathcal{D}$ en introduisant la fonction $g = f + \omega$. La formule de Taylor permet alors d'écrire, à des termes d'ordres supérieurs près (pour la norme $\|.\|_1$), la variation $J(g) - J(f)$ sous la forme :

$$\int_a^b [F'_y(x, f(x), f'(x))\, \omega(x) + F'_{y'}(x, f(x), f'(x))\, \omega'(x)]\, dx.$$

On peut donc dire que la fonctionnelle :

$$\omega \mapsto \int_a^b [F'_y \omega + F'_{y'} \omega']\, dx$$

est la dérivée de la fonction J au point f lorsqu'on munit l'espace $\mathcal{D}$ de la topologie $\mathcal{C}^1$. Avec Lagrange, on notera cette dérivée $\delta J[f]$ et l'on dira qu'elle est la « variation première » de J en f.

On a ainsi démontré qu'une condition nécessaire pour que f soit un minimum relatif faible (et a fortiori fort) de J est que l'on ait $\delta J[f] = 0$.

Équation d'Euler-Lagrange

Si l'on suppose que f est un minimum relatif faible de J deux fois continûment dérivable, on peut transformer l'expression de $\delta J[f]$ en intégrant par partie le second terme. On obtient ainsi :

$$\delta J[f](\omega) = \int_a^b \omega[F'_y - \frac{d}{dx}(F'_{y'})]\, dx.$$

Ce qui conduit à l'équation donnée par Euler en 1744 :

Théorème 1. Une condition ***nécessaire*** pour qu'une fonction f deux fois continûment dérivable soit un minimum relatif faible de J est qu'elle vérifie l'équation :

$$F'_y(x,f(x),f'(x)) - \frac{d}{dx}[F'_{y'}(x,f(x),f'(x))] = 0.$$

Ce résultat est une conséquence immédiate du lemme suivant :

Lemme 1. Soit ***h*** une fonction continue sur ***[a, b].*** Si l'on a :

$$\int_a^b \varepsilon(x)\, h(x)\, dx = 0$$

pour toute fonction ε continûment dérivable sur ***[a,*** b] et vérifiant $\varepsilon(a) = \varepsilon(b) = 0$, on a $h = 0$ sur ***[a, h].***

Démonstration du *lemme 1.* Supposons que la fonction ***h*** soit, par exemple, positive en un point x_0 de]a, b[. On peut alors trouver un intervalle [c, d] contenant x_0 sur lequel h est positive.

Si l'on désigne par ε la fonction égale à $(x \quad c)^2$ (d-x)' sur [c, d] et nulle en dehors de [c, d], on a dans ces conditions :

$$\int_a^b \varepsilon(x)\, h(x)\, dx > 0,$$

qui est en contradiction avec les hypothèses ; par conséquent ***h*** est nulle sur [a, b]. La démonstration est terminée.

L'équation d'Euler-Lagrange peut s'écrire :

$$F'_y - F''_{y'x} - y'F''_{y'y} - y''F''_{y'y'} = 0 ;$$

c'est une équation différentielle du second ordre et sa solution dépend de deux constantes arbitraires, qui sont en général déterminées par les conditions limites en a et en ri. Cette situation diffère donc du problème classique de Cauchy, qui consiste à déterminer une solution d'une équation différentielle du second ordre par sa valeur et celle de sa dérivée en un point.

Exemple *1.* Lorsque la fonction F est indépendante de x, l'équation d'Euler-Lagrange admet l'intégrale première $F - y'F'_{y'} = C^{te}$.

On en déduit par exemple que, pour le problème de la surface minimale de révolution, ses solutions sont, dans les bons cas, les chaînettes :

$$y = \lambda\, \mathrm{ch}\frac{x+\mu}{\lambda}.$$

Remarque 1. On a supposé ici que le minimum f était deux fois continûment dérivable. En fait, une étude plus fine, due à P. Du Bois-Reymond, permet de montrer que, si f est une fois continûment dérivable, la fonction $F'_{y'}(x,f(x),f'(x))$ est dérivable et que sa dérivée est égale à $F'_y(x,f(x),f'(x))$; autrement dit, f satisfait encore l'équation d'Euler-Lagrange.

On peut de plus vérifier qu'elle est alors deux fois dérivable en tout point où l'on a $F''_{y'y'}(x,f(x),f'(x)) \neq 0$. Ainsi l'hypothèse de départ n'était pas trop restrictive.

Remarque 2. Certains problèmes variationnels, par exemple celui qui correspond à la fonction $F = y'^2(1 - y')^2$, n'ont pas de solutions dans l'espace $\mathfrak{D}$. On est ainsi amené à élargir cet espace en l'espace $\mathfrak{D}'$ des fonctions f continues et continûment dérivables par morceaux sur *[a,* b], c'est-à-dire que f est continue sur *[a,* b] et qu'il existe une suite a_0 = a < a, < < a_n = htelleque fsoit continûment dérivable sur chacun des intervalles [a,, *a,+,].*

On peut encore montrer qu'un minimum relatif f de J dans $\mathfrak{D}'$ satisfait l'équation d'Euler-Lagrange sur chacun des intervalles où elle est continûment dérivabie (c'est une conséquence immédiate de la formule de Chasles pour les intégrales) et vérifie de plus les conditions

suivantes, dues à K. Weierstrass et à G. Erdmann : en chaque point de $[a, b]$ les limites à gauche et à droite de :

$$F'_{y'}(x, f(x), f'(x))$$

ainsi que celles de :

$$F'_{y}(x, f(x), f'(x)) - f'(x) F'_{y'}(x, f(x), f'(x))$$

sont égales.

En utilisant le théorème de Rolle, on déduit de cette première condition qu'en un point de discontinuité c def' la fonction $F''_{y'y'}(c, f(c),$ y') à un zéro. Par conséquent, si $F''_{y'y'}(x, y, y')$ est sans zéro, tout minimum de J dans $\mathcal{D}'$ est deux fois continûment dérivable.

Conditions de Legendre et Jacobi

Soit f un minimum relatif faible de J dans $\mathcal{D}$. Utilisant à nouveau la formule de Taylor, on peut écrire, toujours à des termes d'ordres supérieurs près, la variation de J correspondant à une variation ω de f sous la forme :

$$\frac{1}{2}\int_a^b [P(x)\,\omega(x)^2 + 2\,Q(x)\,\omega(x)\,\omega'(x) + R(x)\,\omega'(x)^2]\,dx\,,$$

où l'on a posé :

$$P(x) = F''_{yy}(x, f(x), f'(x)),$$
$$Q(x) = F''_{yy'}(x, f(x), f'(x)),$$
$$R(x) = F''_{y'y'}(x, f(x), f'(x)).$$

La fonctionnelle :

$$\omega \mapsto \int_a^b [P\omega^2 + 2\,Q\omega\omega' + R\omega'^2]\,dx$$

est une forme quadratique sur l'espace $\mathcal{E}$ que l'on peut interpréter comme la dérivée seconde $\delta^2 J[f]$ de J enf: on dira que $\delta^2 J[f]$ est la « variation seconde » de J en f. On a ainsi le résultat suivant : Une condition nécessaire pour quef soit un minimum relatif faible de J est que $\delta^2 J[f]$ soit une forme quadratique positive, c'est-à-dire telle que $\delta^2 J[f]\,(\omega) \geqslant 0$ pour tout $\omega \in \mathcal{E}$.

On va, suivant Legendre, transformer l'expression de cette variation seconde en remarquant que, si w est une fonction continûment dérivable sur [a, b], on a :

$$\int_a^b \frac{d}{dx}(w\,\omega^2)\,dx = 0$$

pour toute fonction $\omega \in \mathcal{E}$. On peut donc écrire :

$$\delta^2 J[f](\omega) = \int_a^b [(P + w')\omega^2 + 2(Q + w)\omega\omega' + R\omega'^2]\,dx$$

soit encore :

$$\delta^2 J[f](\omega) = \int_a^b R\Big[\,\omega' + \frac{Q + w}{R}\,\omega\Big]^2 dx\,,$$

si le discriminant $(Q + w)^2 \quad R(P + w')$ est nul. Cette égalité nous conduit à une seconde condition, énoncée par Legendre en 1786 :

Théorème 2. Une condition nécessaire pour quefsoit un minimum relatif faible de J est que l'on ait sur [a, b] l'inégalité :

$$R(x) = F''_{y'y'}(x, f(x), f'(x)) \geqslant 0.$$

Démonstration. Supposons R négative en un point x_0 de]a, b[. On peut trouver un intervalle $[c, d]$ contenant x_0 sur lequel R reste négative. On peut supposer qu'il existe une solution de l'équation différentielle $(Q + w)^2 - R(P + w') = 0$ sur cet intervalle.

On a alors l'inégalité $\delta^2 J[f](\omega) < 0$ pour la fonction ω qui intervient dans la preuve du lemme 1, ce qui est absurde.

L'expression précédente de $\delta^2 J[f]$ montre que, si l'équation :

$$(Q + w)^2 + R(P + w') = 0$$

a une solution sur [a, b], la condition $R > 0$ entraîne $\delta^2 J[f](\omega) > 0$ pour toute

variation ω non nulle. C'est en partant de cette remarque que Weierstrass montra, en 1879, que les conditions suivantes sont suffisantes pour qu'une fonction f de $\mathfrak{D}$ soit un minimum relatif *faible* de J :

a) La fonction f est une solution de l'équation d'Euler-Lagrange ;

b) On a $F''_{y'y'}(x, f(x), f'(x)) > 0$ sur $[a, b]$;

c) Il existe une solution sur $[a, b]$ de l'équation $(Q + w)^2 - R(P + w') = 0$.

Jacobi introduisit en 1837 le changement de variable $w = -Q - Ru'/u$ qui lui permit de transformer l'équation différentielle $(Q + w)^2 - R(P + w') = 0$ en l'équation différentielle linéaire du second ordre, dite équation *de* Jacobi :

$$(P - Q')u - \frac{d}{dx}(Ru') = 0;$$

et l'existence d'une solution de la première sur $[a, b]$ est équivalente à l'existence d'une solution sans zéro sur $[a, b]$ de la seconde.

Soit alors $u(x)$ une solution non nulle de l'équation de Jacobi telle que $u(a) = 0$. On dira qu'un point $c \geqslant a$ est un *point* conjugué de a si l'on a $u(c) = 0$. Le théorème de Sturm permet de montrer que, si l'intervalle $[a, b]$ ne contient aucun point conjugué de a, il existe une solution de l'équation de Jacobi sans zéro sur cet intervalle, donc une solution du discriminant sur $[a, b]$ (cf. supra).

En 1877, Weierstrass donna la condition suivante, appelée condition de Jacobi :

Théorème 3. Une condition nécessaire pour qu'une fonction f de $\mathfrak{D}$ vérifiant $F''_{y'y'}(x, f(x), f'(x)) > 0$ sur $[a, b]$ soit un minimum relatif faible de J est que l'intervalle $[a, b]$ ne contienne aucun point conjugue du point a.

On remarquera que, contrairement aux théorèmes 1 et 2 qui expriment des conditions locales, la condition de Jacobi est globale.

Remarque 3. Dans son important mémoire de 1837, Jacobi mit aussi en évidence la propriété importante suivante de l'équation qui porte son nom :

Soit $f(x, \lambda)$ une famille différentiable à un paramètre de solutions de l'équation d'Euler-Lagrange telle que $f(x, 0) = f(x)$. Alors la fonction $u(x) = f'_\lambda(x, 0)$ est une solution de l'équation de Jacobi ; autrement dit, l'équation de Jacobi correspondant à la fonction f est l'« équation aux variations » de l'équation d'Euler-Lagrange pour sa solution f.

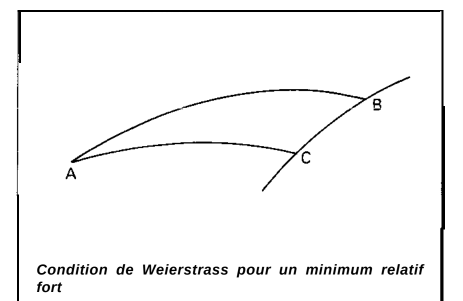
Remarque 4. Les conditions données dans les théorèmes 1, 2 et 3 concernent un minimum relatif faible. En 1879, Weierstrass donna la condition nécessaire suivante pour un minimum relatif fort, obtenue en exprimant que la valeur de J sur l'arc AB doit être inférieure à la somme des valeurs de J sur les arcs AC et CB (cf. figure). L'expression :

$$F(x, f(x), y') - F(x, f(x), f'(x)) - (y' - f'(x))F'_{y'}(x, f(x), y')$$

doit être positive pour tout y' et tout x dans $[a, b]$; cette condition est en particulier vérifiée si l'on a :

$$F''_{y'y'}(x, f(x), y') > 0$$

pour tout y' et tout x dans $[a, b]$.



Condition de Weierstrass pour un minimum relatif fort

Il montra également que les conditions suivantes sont suffisantes pour qu'une fonctionfde 𝔇 soit un minimum relatif *fort* de J :

a) La fonction f est solution de l'équation d'Euler-Lagrange ;

b) On a l'inégalité $F''_{y'y'}(x, y, y') > 0$ pour tout y' et pour tout couple (x, y) dans un voisinage du graphe de f;

c) L'intervalle [a, *b*] ne contient aucun point conjugué du point *a*.

Théorie de Morse

Dans son aspect classique, la théorie de Morse ne fait pas partie du calcul des variations. Elle concerne en fait l'étude des fonctions différentiables sur les variétés et permet, en particulier, de donner des décompositions des variétés jouant en topologie différentielle le rôle que jouent les décompositions simpliciales en topologie combinatoire. C'est en utilisant cette technique que S. Smale démontra, en 1962, la conjecture de Poincaré en dimensions supérieures à 5.

Si f est une fonction différentiable à valeurs réelles sur une variété M, on dit qu'un point z de M est un *point critique* de f s'il annule sa différentielle df, ce qui s'exprime, dans un système de coordonnées locales $x_1, \dots x_n$ tel que $x_i(z) = 0$, par les conditions :

$$\frac{\partial f}{\partial x_1}(0) = \cdot = \frac{\partial f}{\partial x_n}(0) = 0.$$

Ce point critique est non *dégénéré* si le hessien $H(f)$ de f en z, c'est-à-dire la forme quadratique définie par la matrice :

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(0)\right)$$

est de rang maximum. *L'index* de z est alors le nombre de valeurs propres négatives du hessien.

Le lemme de Morse assure que, si z est un point critique non dégénéré d'index p, il existe un système de coordonnées locales $y_1, \dots, y_n$ avec $y_i(z) = 0$ tel que l'on ait :

$$(y_1, \dots yn) = f(0) - y_1^2 - \dots - y_p^2 + y_{p+1}^2 + \quad + .Y:;$$

ce qui montre en particulier que les points critiques non dégénérés sont isolés. On peut déduire de cette expression que, si a et b ne sont pas des valeurs critiques de f, et si l'intervalle $]a, b[$ contient une seule valeur critique, correspondant à un seul point critique z, non dégénéré, on obtient la sous-variété $M_b = f^{-1}(]-\infty, b])$ en recollant à la sous-variété $M_a = f^{-1}(]-\infty, a])$ une « anse » $D^p \times D^{n-p}$ d'indice p, où p est l'index de z, au moyen d'une application différentiable de $S^{p-1} \times D^{n-p}$ dans le bord de M,,. La variété M_b a donc le type d'homotopie de l'espace obtenu en recollant à M, une cellule de dimension p. On en déduit par exemple que le q-ième nombre de Betti de M est inférieur au nombre de points critiques d'index q de f.

L'originalité de Morse fut alors de montrer, en 1934, que, sur une variété riemannienne M, on pouvait raisonner de façon analogue pour l'espace $\Omega(M ; p, q) = \Omega$ des courbes $\mathcal{C}^\infty$ par morceaux joignant p à q (qui est une variété banachique) et la fonction $E : \Omega \rightarrow R$ définie par :

$$CU \int \left\| \frac{dc}{dt} \right\|^2 dt,$$

qui est différentiable sur Ω.

On se trouve ici devant un problème variationnel pour lequel les solutions de l'équation d'Euler-Lagrange, qui sont les points critiques de E, sont les géodésiques $\mathcal{C}^\infty$ joignantp à q (cf. *Remarques 1* et *2)*. Le hessien de E pour une géodésique y est la variation seconde de E en y, et y est un

point critique non dégénéré de E si et seulement si cette variation seconde est une forme définie. On peut montrer qu'il en est ainsi si le point q n'est pas conjugué du point p (en un sens qui généralise directement celui du chapitre 5) le long de γ.

Dans ces conditions, le plus grand sous-espace sur lequel le hesrien est défini négatif est de dimension finie, et sa dimension est égale au nombre de points conjugués de p sur y comptés avec leur ordre de multiplicité (puisque l'équation de Jacobi est dans ce cas une équation linéaire du second ordre et de dimension n, cette multiplicité est toujours inférieure à n).

Le résultat central de la théorie de Morse assure alors que, si p et q ne sont conjugués le long d'aucune géodésique, l'espace Ω a le type d'homotopie d'un complexe simplicial dénombrable ayant une cellule de dimension d pour chaque géodésique d'index djoignantp à q (en fait, ce type d'homotopie est un invariant topologique de la variété).

Par exemple, les géodésiques de la sphère euclidienne sont les arcs de grands cercles, et deux points sont conjugués si et seulement s'ils sont diamétralement opposés ; la multiplicité pour le demi-grand cercle est alors $n - 1$. On en déduit que, si p et q ne sont pas diamétralement opposés, l'espace $\Omega(\mathbf{S}^n ; p, q)$ a le type d'homotopie d'un complexe simpiicial ayant une cellule en dimensions 0, $n - 1$, $2(n-1)$, $3(n \quad 1)$,

Ce résultat, qui reste vrai pour toute variété riemannienne homéomorphe à $\mathbf{S}^n$, permet de montrer l'existence globale de géodésiques.

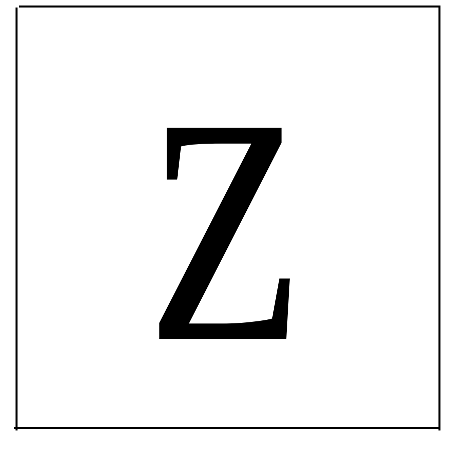
En 1957, R. Bott, appliquant ces méthodes aux groupes unitaire U(n, C) et orthogonal O(n, R), a montré, ce qui est un résultat fondamental pour la topologie moderne, la périodicité des groupes d'homotopie de ces espaces : par exemple, $\pi_{2i}(\mathbf{U}) = 0$ et $\pi_{2i+1}(\mathbf{U}) = \mathbf{Z}$ pour le groupe unitaire

$$\mathbf{U} = \varinjlim \mathbf{U}(n, \mathbf{C}).$$

CLAUDE GODBILLON

Bibliographie

N. I. AKHIEZER, *Calculus of Variations*, Gordon and Breach Science, New York, 1988 / J.-P. BOURGUIGNON, *Calcul variationnel*, *École* polytechnique, Palaiseau, 1993 / I. M. GELFAND & S. V. FOMIN, *Calculus of Variations*, Prentice Hall, Englewood Cliffs (N.J.), 1963 / M. R. HESTENES, *Calculus of Variation and Optimal Control Theory*, R. E. Krieger Publ., New York, 1980 / J. MILNOR, *Morse Theory*, Princeton Univ. Press, Princeton (N.J.), 1963 / M. MORSE, *The Calculus of Variations in the Large*. A.M.S., New York, 1934. rééd. American Mathematical Society, Providence (R.I.), 1986 / L. S. PONTRIAGUINE, L. BOLTIANSKI & V. GAMKRELIDZE, *Théorie mathématique des processus optimaux*, M.I.R., Moscou. 1978.



ZÊTA FONCTION

Issues d'un calcul formel d'Euler, la « fonction zêta » de Riemann et les « fonctions L » de Dirichlet ont été jusqu'ici les outils analytiques les plus puissants pour étudier la répartition et les propriétés des nombres premiers. Mais ces fonctions sont elles-mêmes devenues l'objet d'études analytiques poussées, en raison de leurs propriétés très particulières qui semblent être liées aux comportements les plus cachés de la théorie des nombres et sont encore loin d'être bien comprises.

Le mouvement d'idées qui tend, depuis 1920, à l'unification de la théorie des nombres et de la géométrie algébrique a conduit à définir, dans cette dernière théorie, des « fonctions zêta » et des « fonctions L » analogues aux fonctions classiques et présentant un comportement semblable. Il y a lieu de penser qu'on se trouve en présence de fragments encore mal reliés d'une vaste théorie générale, participant de l'analyse, de la théorie des groupes et de la géométrie algébrique, qui nous fera un jour pénétrer dans les recoins les plus mystérieux de la « reine des mathématiques » (C. F. Gauss), l'étude des nombres entiers.

La fonction zêta de Riemann

La série :

$$(1)\qquad \sum_{n=1}^{\infty} n^{-s},$$

avec $n^{-s} = \exp(-s \ln n)$, et le produit infini :

$$(2)\qquad \prod_{p} (1-p^{-s})^{-1},$$

étendu aux nombres premiers p, sont tous deux absolument convergents pour $s = \sigma + it$ de partie réelle $\sigma > 1$ et représentent la même fonction analytique $\zeta(s)$ dans ce domaine. Le résultat fondamental de Riemann est qu'il est possible de prolonger cette fonction en une fonction ***mèromorphe*** dans tout le plan, vérifiant l'équation fonctionnelle :

$$(3)\qquad \xi(s) = \xi(1-s),$$

où l'on a posé :

$$(4)\qquad \xi(s) = \frac{1}{2}s(s-1)\pi^{-s/2}\Gamma\left(\frac{s}{2}\right)\zeta(s).$$

Une des démonstrations de Riemann lie la fonction zêta à une fonction thêta de Jacobi, grâce à l'expression de $\Gamma(s)$ par l'intégrale eulérienne qui donne :

$$\pi^{-s}\Gamma(s)\zeta(2s) = \int_0^{\infty} t^{s-1}\left(\theta(it) - \frac{1}{2}\right)dt,$$

où l'on a :

$$(5)\quad \theta(x) = \frac{1}{2}\sum_{n=-\infty}^{\infty} \exp(\pi i n^2 x)$$

$$\frac{1}{2} + \sum_{n=1}^{\infty} \exp(\pi i n^2 x),$$

pour Imx > 0. L'identité fondamentale qui exprime la propriété « modulaire » de la fonction thêta :

$$(6)\qquad \theta\left(-\frac{1}{x}\right) = (-ix)^{1/2}\,\theta(x),$$

avec la détermination de la racine carrée $z^{1/2}$ positive pour z réel > 0, donne alors :

$$(7)\quad \pi^{-s}\Gamma(s)\zeta(2s) = \int_1^{\infty} (t^{s-1} + t^{-s-1/2})\left(\theta(it) - \frac{1}{2}\right) dt$$

$$\frac{1}{2s}\quad \frac{1}{2s-1}$$

Vu la décroissance exponentielle de $\theta(it)$ $1/2$ à l'infini, l'intégrale dans cette formule converge pour *toutes* les valeurs complexes de s et ne change pas quand on remplace s par $-s + 1/2$, d'où l'équation (3).

On voit aussitôt que $s = 1$ est le seul pôle de $\zeta(s)$; il est simple et de résidu 1 ; les points 2, -4, sont des zéros simples de $\zeta(s)$, dits « triviaux ». Les seuls autres zéros de $\zeta(s)$ sont tels que $0 \leqslant \sigma \leqslant 1$ et Riemann a émis l'hypothèse, non encore démontrée, que tous ces zéros sont sur la droite $\sigma = 1/2$. On a en outre, pour le nombre N(T) des zéros contenus dans le rectangle $0 \leqslant \sigma \leqslant 1, 0 \leqslant t \leqslant T$, l'expression asymptotique :

$$(8)\quad N(T) = \frac{1}{2\pi} T \ln T - \frac{1 + \ln 2\pi}{2\pi} T + O(\ln T).$$

où 0 désigne le symbole de Landau (cf. calculs ASYMPTOTIQUES, chap. 1).

Riemann avait annoncé sans démonstration qu'il y a en fait une infinité de zéros sur la droite $\sigma = 1/2$; ce résultat fut prouvé par G. H. Hardy en 1914 et A. Selberg a établi en 1942 qu'il y aune constante A > 0 telle que le nombre de zéros pour lesquels on a $\sigma = 1/2$ et $0 \leqslant t \leqslant T$ est supérieur ou égal à AT ln T. On a calculé numériquement plus de trois millions de zéros de $\zeta(s)$ et on les a tous trouvés sur la droite $\sigma = 1/2$. Il faut noter toutefois qu'on connaît des exemples de séries de Dirichlet vérifiant des équations fonctionnelles du type (3) et ayant cependant une infinité de zéros où CI > 1.

Fonction zêta et fonctions L d'un corps de nombres algébriques

R. Dedekind généralisa la définition des fonctions zêta et L à un corps de nombres algébriques k, en prenant :

$$(9)\quad L(s, \chi) = \sum_{\mathfrak{a}} \chi(\mathfrak{a})(N\mathfrak{a})^{-s}$$

$$= \prod_{\mathfrak{p}} (1 - \chi(\mathfrak{p})(N\mathfrak{p})^{-s})^{-1},$$

où $\mathfrak{a}$ parcourt l'ensemble des idéaux entiers de k, où p parcourt l'ensemble des idéaux premiers, où Na est la norme de l'idéal a, c'est-à-dire le nombre d'éléments de $\mathfrak{o}/\mathfrak{a}$ (où o est l'anneau des entiers de k) et où χ est un caractère du groupe des idéaux $\neq 0$ (pour $\chi = 1$, on a la fonction zêta). E. Ecke put établir que ces fonctions sont méromorphes et vérifient des équations fonctionnelles analogues à (3). Il introduisit d'autres fonctions L à l'aide de caractères plus généraux que ceux de Dedekind et put encore, au prix de calculs difficiles, prouver l'existence d'équations fonctionnelles. Dans sa thèse de 1950, J. Tate a montré comment la théorie des idèles

permet une exposition simple et unifiée de tous ces travaux qui, en définitive, peuvent être fondés sur la transformation de Fourier dans le groupe des adèles de k (rappelons que la propriété (6) est une conséquence de la formule de Poisson de la théorie de Fourier classique). Dans la théorie de Tate, on considère d'une part le groupe additif k_A des adèles de k et sa mesure de Haar $d\mathbf{a}$, normalisée de sorte que k_A/k ait pour mesure 1 ; il y a sur k_A un caractère $\lambda : k_A \to T = \mathbf{R}/\mathbf{Z}$ tel que l'application :

$$(\mathbf{a}, \mathbf{b}) \mapsto \exp(2\pi i \lambda(\mathbf{ab}))$$

identifie k_A à son dual de Pontriaguine. La transformée de Fourier d'une fonction intégrable sur k_A est alors donnée par :

$$(\mathcal{F}f)(\mathbf{b}) = \int f(\mathbf{a}) \exp(-2\pi i \lambda(\mathbf{ab}))\, d\mathbf{a}$$

et on a la formule d'inversion habituelle $(\mathcal{F}(\mathcal{F}f))(\mathbf{a}) = f(-\mathbf{a})$ si f est suffisamment régulière. On considère d'autre part le groupe multiplicatif des idèles k_A^* et sa mesure de Haar $d^*\mathbf{a}$. On appelle *quasi-caractère* de k_A^* un homomorphisme continu $c : k_A^* \to \mathbf{C}^*$ qui est *égal à* 1 *sur* k^* (de sorte qu'il s'agit en fait d'un homomorphisme dans $\mathbf{C}^*$ du groupe k_A^*/k^* des *classes d'idèles).* Pour un tel quasi-caractère c, il y a un nombre réel μ bien déterminé tel que l'on ait $|c(\mathbf{a})| = |\mathbf{a}|^{\mu}$ pour tout idèle $\mathbf{a}$; on dit que c'est l'*exposant* de c. Pour des «fonctions poids » $f : k_A \to \mathbf{C}$ satisfaisant à certaines conditions de régularité, on pose alors :

$$(10) \qquad \zeta(f, c) = \int f(\mathbf{a}) c(\mathbf{a})\, d^*\mathbf{a},$$

intégrale qui a un sens pour tout quasi-caractère d'exposant > 1.

Un quasi-caractère peut toujours s'écrire (de plusieurs manières) $c(\mathbf{a}) = \chi(\mathbf{a})\,|\mathbf{a}|^s$, où χ est un *caractère* de k_A^*/k^* (« Grossencharakter » dans la terminologie de Hecke) et s un nombre complexe ; deux quasi-caractères correspondant au même caractère χ sont dits équivalents ; l'ensemble des quasi-caractères équivalents à un quasi-caractère donné a donc une structure complexe et l'on peut parler de « prolongement analytique » d'une fonction holomorphe dans un ensemble ouvert de l'ensemble des quasi-caractères.

Le théorème fondamental de Tate est alors que, pour une fonction poids f donnée la fonction $c \mapsto \zeta(f, c)$, définie seulement pour les quasi-caractères d'exposant > 1, se prolonge en une fonction *méromorphe* sur l'ensemble de tous les quasi-caractères ; ses seuls pôles sont le caractère trivial $\chi_0 : \mathbf{a} \mapsto 1$ et le quasi-caractère « module » $N : \mathbf{a} \mapsto |\mathbf{a}|$, avec des résidus respectivement égaux à $\beta \cdot f(0)$ et $-\beta \cdot \mathcal{F}f(0)$, où β est le volume de k_A^*/k^* pour la mesure de Haar additive $d\mathbf{a}$. Enfin, on a l'équation fonctionnelle :

$$(11) \qquad \zeta(f, c) = \zeta(\mathcal{F}f, \hat{c}),$$

où $\hat{c}$ est le quasi-caractère $\hat{c}(\mathbf{a}) = |\mathbf{a}|\,(c(\mathbf{a}))^{-1}$ (de sorte que $\hat{\hat{c}} = c$). La démonstration est une adaptation facile de celle de Riemann rappelée plus haut ; on décompose l'intégrale (10) en deux autres, étendues respectivement aux idèles tels que $|\mathbf{a}| \leqslant 1$ et aux idèles tels que $|\mathbf{a}| \geqslant 1$, et on ramène la première au domaine $|\mathbf{a}| \geqslant 1$ par changement de variable et utilisation de la formule générale de Poisson de l'analyse harmonique.

Pour retrouver à partir du théorème de Tate les résultats de Hecke et de Dedekind, on spécialise la fonction poids ; on prend :

$$f(\mathbf{a}) = \prod_v f_v(\mathbf{a}_v),$$

où, pour chaque place v de k, a est le composant de l'idèle a dans le corps local k_v et f_v une fonction convenable sur k D'autre part, si l'on a c(a) = $\chi(\mathbf{a})$ a ^{s}, on désigne par S_χ l'ensemble fini des places de k formé des places infinies et des places finies où χ prend des valeurs $\neq 1$ dans le groupe des unités de k Alors, si l'on pose :

$$(12) \quad L(s, \chi) = \prod_{\mathfrak{p} \notin S_\chi} (1 - \chi(\mathfrak{p})(N\mathfrak{p})^{-s})^{-1},$$

on trouve l'égalité $\zeta(f,$ c) = $\Phi(s)L(s, \chi)$ où $\Phi(s)$ est un produit de fonctions de la forme $s \mapsto \Gamma(\alpha s + \beta)^m$, expression dans laquelle α et β sont des constantes complexes et m un entier rationnel, et de fonctions rationnelles par rapport à un certain nombre de puissances ρ^s, où les ρ sont des constantes non nulles. L'équation fonctionnelle (11) se réduit alors à l'équation de Hecke pour la fonction L.

Fonctions zêta et fonctions L sur une variété algébrique définie sur un corps fini

Depuis les travaux de E. Artin, on sait que tous les résultats de la théorie des nombres algébriques se transportent (avec des expressions plus simples, dues à l'absence des « places infinies ») aux «corps de fonctions algébriques d'une variable sur un corps fini F, », c'est-à-dire les extensions algébriques finies du corps des fractions rationnelles F,(X). E. Artin lui-même avait noté, sur le cas particulier des extensions quadratiques de F,(X), que la définition de Dedekind de la fonction zêta se généralise à un tel corps k en prenant pour $\mathfrak{p}$ toutes les places de k et pour $N\mathfrak{p}$ le nombre d'éléments du corps résiduel de la place $\mathfrak{p}$. La théorie de Tate s'étend également sans difficulté.

Mais on peut considérer k comme le corps des fonctions rationnelles sur une *courbe algébrique* irréductible définie sur $\mathbf{F}_q$; ce point de vue amène à une nouvelle généralisation, en remplaçant la courbe par une *variété algébrique* X de dimension quelconque définie sur $\mathbf{F}_q$. Pour simplifier, on supposera qu'il s'agit d'une variété affine, ensemble des points x = $(x_1, \ldots, x_m)$ d'un espace $\overline{\mathbf{F}}_q^m$, où $\overline{\mathbf{F}}_q$ est la clôture algébrique de $\mathbf{F}_q$, vérifiant un nombre fini d'équations $P_\alpha(x_1, \ldots$ x,,) = 0, où les P_α sont des polynômes à coefficients dans F,,. Soit a l'idéal de l'anneau de polynômes $\mathbf{F}_q[T_1, \ldots$ T,,] engendré par les P,. Tout point x $\in$ X définit un homomorphisme $\mathbf{F}_q[T_1, \ldots, T_m] \rightarrow \overline{\mathbf{F}}_q$ transformant T, en x_j pour $1 \leqslant j \leqslant m$ et s'annulant dans a ; réciproquement, un tel homomorphisme correspond à un point x $\in$ X et à un seul. L'image de cet homomorphisme est un corps fini, extension de $\mathbf{F}_q$, ayant donc q^h éléments ; on pose $h = \deg(x)$. On montre alors que le produit infini :

$$(13) \quad Z(X, t) = \prod_{x \in X} (1 - t^{\deg(x)})$$

est, pour $t < q^{-\dim(x)}$, absolument convergent et l'on définit la fonction zêta de X par :

$$(14) \quad \zeta(X, s) = Z(X, q^{-s}).$$

On voit facilement que, pour tout entier $n \geqslant 0$, il n'y a qu'un nombre fini v,, de points de X tels que $\deg(x) = n$, et on déduit de (13) l'égalité :

$$(15) \quad \operatorname{Log} Z(X,t) = \sum_{n=1}^{\infty} \nu_n \frac{t^n}{n}$$

Le nombre ν_n s'interprète à l'aide de *l'uutomorphisme de Frobenius* F de X, qui à tout point $(x_1, \ldots, x_m)$ de X fait correspondre le point $(x_1^q, \ldots, x_m^q)$; ν_n est

simplement le nombre des points de X *fixes par* F^n. Cette interprétation, d'abord introduite par A. Weil, est à la base de tous les résultats récents obtenus sur les fonctions zêta des variétés X.

On définit de la même manière la fonction $Z(X, t)$ lorsque X est une variété projective sur F, ou une variété « abstraite » au sens de A. Weil ou de J.-P. Serre. Lorsque X est une courbe projective sans singularité de genre g, F. K. Schmidt a montré en 1929 que l'on peut écrire :

$$(16) \qquad Z(X, t) = \frac{P_{2g}(t)}{(1-t)(1-qt)},$$

où P_{2g} est un polynôme de degré $2g$, et on a l'équation fonctionnelle :

$$(17) \qquad Z\left(\frac{1}{qt}\right) = q^{1-g} t^{2-2g} Z(t);$$

en outre, H. Hasse pour $g = 1$ et A. Weil pour le cas général montrèrent que les zéros de P_{2g} sont tels que $|t| = q^{1/2}$, ce qui correspond dans ce cas à l'« hypothèse de Riemann ». Pour X (projective ou non) de dimension quelconque, B. Dwork montra en 1960 que $Z(X, t)$ est encore une fonction *rationnelle* de t. Par exemple, si $X = \overline{F}_q^m$, on a $\zeta(X, s) = (1 - q^{m-s})^{-1}$. Grâce à l'introduction d'une notion de « cohomologie » pour les variétés sur un corps quelconque, A. Grothendieck et M. Artin ont montré que, si X est une variété projective irréductible sans singularité de dimension n sur F_q, la fonction zêta vérifie l'équation fonctionnelle généralisant (17) :

$$(18) \qquad Z\left(\frac{1}{q^n t}\right) = (-1)^k q^{nk/2} t^k Z(t),$$

où k est la « caractéristique d'Euler-Poincaré » de X pour cette cohomologie. Mais on n'a pas encore obtenu de démonstration de l'« hypothèse de Riemann » correspondante qui serait que les zéros de $Z(X, t)$ soient tous sur le cercle $|t| = q^{n/2}$.

Suivant une idée de E. Artin, on peut aussi définir des « fonctions L » relatives à l'action d'un groupe fini G (commutatif ou non) opérant dans la variété X ; on pose :

$$(19) \qquad \ln L(X, \chi; t) = \sum_{n=1}^{\infty} \nu_n(\chi) \frac{t^n}{n}$$

pour un *caractère* χ de G, les nombres $\nu_n(\chi)$ étant définis par :

$$(20) \quad \nu_n(\chi) = (\text{Card } G)^{-1} \sum_{s \in G} \chi(s^{-1}) \Lambda(sF_n),$$

où $\Lambda(sF^n)$ est le nombre de points de X fixes par l'automorphisme sF^n. A. Grothendieck et J.-L. Verdier ont montré que ces fonctions sont encore *rationnelles*. Une des propriétés les plus importantes de ces fonctions est qu'elles fournissent une *factorisation* de la fonction zèta :

$$(21) \qquad Z(X, t) = \prod_{\chi} L(X, \chi; t)^{\deg(\chi)},$$

où χ parcourt l'ensemble des caractères de G.

Fonction zêta et fonctions L sur une variété algébrique « définie sur Z »

Considérons maintenant dans l'anneau de polynômes $Z[T_1, ..., T_m]$ un idéal a et convenons de dire qu'il définit une « variété X sur Z » (le langage adapté à cette situation est celui des « schémas » de Grothendieck). Pour chaque nombre premier p, l'homomorphisme canonique :

$$\mathbf{Z} \rightarrow \mathbf{Z}/p\mathbf{Z} = \mathbf{F}_p$$

définit un homomorphisme :

$$\mathbf{Z}[T_1, ..., T_m] \rightarrow \mathbf{F}_p[T_1, ..., T_m]$$

transformant a en un idéal a,,, définissant par suite une variété algébrique X_p sur $\mathbf{F}_p$. On peut alors, tout au moins formellement, considérer le produit :

$$(22) \qquad \zeta(X, s) = \prod \zeta(X_p, s)$$

étendu à tous les nombres premiersp ; c'est la *fonction zêta de Hase- Weil.* Par exemple, si l'on prend a = (0), on trouve $\zeta(X, s) = \zeta(m \quad s)$ ou, au second membre, ζ est la fonction de Riemann. Si n est la dimension de X (qu'on définit ici comme le plus grand nombre tel qu'il y ait une chaîne strictement croissante $\mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \subset \mathfrak{p}_{n+1}$ d'idéaux premiers de $\mathbf{Z}[T_1, \dots T_n]$ contenant a), on montre que le produit (22) est absolument convergent pour Re $s > n$. On *conjecture* que $\zeta(X, s)$ peut se prolonger en une fonction méromorphe dans tout le plan et vérifiant une équation fonctionnelle analogue à (3) : mais on ne sait jusqu'ici prouver cette conjecture que dans un petit nombre de cas où X est soit une courbe algébrique d'un type très particulier, soit une variété abélienne d'un type spécial, soit enfin certaines variétés fibrées ayant pour base une courbe algébrique et pour fibres des variétés abéliennes. Dans chacun de ces cas, on parvient au résultat par un calcul explicite de $\zeta(X, s)$ à l'aide de fonctions L de Hecke ou de Dedekind. En général, on sait seulement prolonger analytiquement $\zeta(X, s)$ en une fonction méromorphe dans le demi-plan Re $s > n - 1/2$ et, si X $\neq \emptyset$, le point $s = n$ est un pôle d'ordre égal au nombre de composantes irréductibles de X de dimension n.

On peut aussi considérer les fonctions L que l'on définit par une formule analogue à (22) :

$$(23) \qquad L(X, \chi; s) = \prod_p L(X_p, \chi; s)$$

lorsque G opère sur X, c'est-à-dire lorsque G opère sur $\mathbf{Z}[T_1, \dots T_n]$ en laissant stable l'idéal a ; on a donc encore une factorisation :

$$(24) \qquad \zeta(X, s) = \prod L(X, \chi; s)^{\deg(\chi)}.$$

Équations fonctionnelles et représentation des groupes

On peut considérer que l'intégrale eulérienne :

$$\Gamma(s) = \int_0^\infty x^{s-1} e^{-x}\, dx$$

définit Γ comme « transformée de Mellin » de e^{-x}, la transformation de Mellin se déduisant de la transformation de Laplace bilatère (ou transformation de Fourier-Laplace) qui à une fonction f fait correspondre la fonction :

$$s \mapsto \int_{-\infty}^{+\infty} e^{ts} f(t)\, dt,$$

par le changement de variable $x = e^t$ dans l'intégrale ; la « formule d'inversion » de la transformation de Mellin donne alors :

$$e^{-x} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-s}\, \Gamma(s)\, ds,$$

où l'intégrale est prise le long de la droite $t \mapsto c + it$ dans le plan complexe, avec $c > 0$. La démonstration de l'équation fonctionnelle de $\zeta(s)$ par Riemann rappelée plus haut conduit à l'idée plus générale d'attacher à une série de Dirichlet :

$$(25) \qquad \phi(s) = \sum_{n=1} a_n n^{-s},$$

convergente dans un demi-plan, donc telle que $a_n = O(n^c)$ pour $c > 0$, la fonction f, analytique dans le demi-plan H tel que Im $z > 0$, définie par :

$$(26) \qquad f(z) = \sum^{\infty} a_n \exp\left(2\pi i n \frac{z}{\lambda}\right),$$

pour $\lambda > 0$, de sorte que la fonction :

$$(27) \qquad \Phi(s) = \left(\frac{2\pi}{\lambda}\right)^{-s} \Gamma(s)\,\phi(s)$$

soit transformée de Mellin de f. Généralisant la méthode de Riemann, E. Hecke a remarqué que les propriétés (A) et (B) suivantes sont équivalentes.

Propriété (A). La fonction :

$$\Phi(s) + \frac{a_0}{s} + \varepsilon\frac{a_0}{k - s},$$

ou $\varepsilon = \pm 1$, est entière et bornée dans chaque « bande » Re $s \leqslant a$ et vérifie, pour un $k > 0$, l'équation fonctionnelle :

$$(28) \qquad \Phi(k - s) = \varepsilon\Phi(s).$$

Propriété (B). On a :

$$f\left(-\frac{1}{z}\right) = \varepsilon\left(\frac{z}{i}\right)^{k} f(z).$$

Lorsque $\Phi(s) = \zeta(2s)$, on a 6 = 1, $\lambda = 2$ et $k = 1/2$. Une fonction analytique dans H de la forme (26), donc telle que $f(z + A) = f(z)$, et vérifiant en outre (B) est dite *forme modulaire de dimension k et de multiplicateur* ε pour le groupe G(h) engendré par les deux automorphismes $z \mapsto z + A$ et $z \mapsto -1/z$ du demi-plan H. Le cas le plus important est celui où A = 1, le groupe G(1) n'étant autre alors que le *groupe modulaire* des transformations :

$$z \mapsto \frac{az + b}{cz + d},$$

avec a, b, c et d dans Z et avec ad – hc = 1 ; ce groupe est le quotient de SL(2, Z) par son centre.

On aperçoit donc là le début d'une étroite relation entre la théorie des représentations des groupes (du groupe GL(2) pour commencer) et les propriétés arithmétiques des courbes algébriques définies sur un corps de nombres, par le biais des fonctions zêta et L attachées à ces courbes.

JEAN DIEUDONNÉ

Bibliographie

T. M. APOSTOL, *Modular Functions und Dirichlet Series in Number Theory*, Springer, New York, 2' éd. 1989 / H. M. EDWARDS, *Riemann's Zeta Function*, Acad. Press, New York-Londres, 1974 / H. JACQUET & R. LANGLANDS, *Automorphic Forms on* GL(2), Springer, Heidelberg-New York, vol. 1, 1970, vol. II, 1972 / S. LANG, *Algebraic Number Theory*, 1970, rééd. Springer, 1986 / G. SHIMURA, *Introduction to the Arithmetic Theory of Automorphic Functions*, Princeton Univ. Press, Princeton (N. J.), 1971 / E. C. TITCHMARSH, *The Theory of the Riemann Zêta-Function*, Oxford Univ. Press, Oxford, 2' éd. 1987 / A. WEIL, *Dirichlet Series and Automorphic Forms*, Springer, New York-Berlin-Heidelberg, 1971.

INDEX

Certaines ENTRÉES de cet index sont aussi des titres d'articles : dans ce cas, elles sont précédées d'une puce et suivies d'un folio (exemple : • ALGÈBRE 12).
En l'absence de puce et de folio, le mot joue seulement le rôle d'entrée d'index (exemple : ADÈLES).
Pour plus d'informations sur le fonctionnement de l'index, voir page 9.

TABLE DES AUTEURS

TABLE DES AUTEURS

Le présent volume a été achevé d'imprimer
sur les presses de l'imprimerie Maury à Manchecourt
en juillet 1997.

Imprimé en France

Dépôt légal septembre 1997
N° d'éditeur : 16760
N° d'imprimeur. 57621M
I.S.B N. E-226-09423-7